

Local Rademacher Complexity-based Learning Guarantees for Multi-Task Learning

Niloofer Yousefi¹, Yunwen Lei², Marius Kloft³
Mansoor Mollaghasemi⁴ and Georgios Anagnostopoulos⁵

¹Department of Computer Science, University of Central Florida

²Department of Mathematics, City University of Hong Kong

³Department of Computer Science, Humboldt University of Berlin

⁴Department of Industrial Engineering, University of Central Florida

⁵Department of Electrical and Computer Engineering, Florida Institute of Technology

niloofaryousefi@knights.ucf.edu, yunwen.lei@hotmail.com, kloft@hu-berlin.de,
Mansoor.Mollaghasemi@ucf.edu and georgio@fit.edu

Abstract

We show a Talagrand-type of concentration inequality for MTL, using which we establish sharp excess risk bounds for Multi-Task Learning (MTL) in terms of distribution- and data-dependent versions of the Local Rademacher Complexity (LRC). We also give a new bound on the LRC for strongly convex hypothesis classes, which applies not only to MTL but also to the standard i.i.d. setting. Combining both results, one can now easily derive fast-rate bounds on the excess risk for many prominent MTL methods, including—as we demonstrate—Schatten-norm, group-norm, and graph-regularized MTL. The derived bounds reflect a relationship akin to a conservation law of asymptotic convergence rates. This very relationship allows for trading off slower rates w.r.t. the number of tasks for faster rates with respect to the number of available samples per task, when compared to the rates obtained via a traditional, global Rademacher analysis.

Keywords: Multi-task Learning, Kernel Methods, Generalization Bound, Local Rademacher Complexity

1 Introduction

Multi-Task Learning (MTL) refers to the concurrent learning of a collection of conceptually related tasks, each of which features its own data. Such an approach can be advantageous over learning each task independently, when the tasks lack a sufficient body of observed data. MTL is accomplished by jointly constraining the tasks' hypothesis spaces, so that tasks mutually regularize the learning of others, based on their inter-task relatedness; this exchange mechanism is often referred to as *information sharing*. Pioneering works on MTL include [16, 9, 2, 4]. Nowadays, MTL frameworks are routinely employed in a variety of settings. Some recent examples include image segmentation [1], HIV therapy screening [12], collaborative filtering [15], age estimation from facial images [55], and sub-cellular location prediction [53] just to name a few prominent ones.

MTL learning guarantees centered around the notion of (global) Rademacher averages and associated complexities, notions that were put forward in [25] and further developed in [8], have been notably pursued in [38], [39], [21], [41], [40] and [42]; these works are briefly surveyed in Sect. 1.3. Let T denote the number of tasks being co-learned and n denote the number of available observations per task. Then, in terms of convergence rates w.r.t. n and T , the fastest-converging error or excess risk bounds derived in these works, whether distribution- or data-dependent, are of the order $O(1/\sqrt{nT})$.

More recently, the seminal works in [26] and [7] introduced a more nuanced variant of these complexities, termed Local Rademacher Complexity (LRC) (as opposed to the original Global Rademacher Complexity (GRC)). This new, modified function class complexity measure is attention-worthy, since, as shown in [7], a LRCs-based (local) analysis is capable of producing more rapidly-converging excess risk bounds, when compared to the ones obtained via a GRC (global) analysis. This can be attributed to the fact that, unlike LRCs, GRCs ignore the fact that learning algorithms typically choose well-performing hypotheses that belong only to a subset of the entire hypothesis space under consideration. The end result of this distinction empowers a local analysis to provide less conservative and, hence, sharper bounds than when a global analysis is employed. To date, there have been only a few additional works attempting to reap the benefits of such local analysis in various contexts: active learning for binary classification tasks [27], multiple kernel learning [23] and [17], transductive learning [52], semi-supervised [44] and bounds on the LRCs via covering numbers [30].

1.1 Our Contributions

Through one of Bousquet’s Talagrand-type concentration inequalities adapted to the MTL case, this paper’s main contribution is the derivation of sharp bounds on the excess MTL risk in terms of the distribution- and data-dependent LRC. For a given number of tasks T , these bounds admit faster (asymptotic) convergence characteristics in the number of observations per task n , when compared to corresponding bounds hinging on the GRC. Thence, these faster rates allow for heightened confidence that the MTL hypothesis selected by a learning algorithm approaches the best-in-class solution as n increases beyond a certain threshold. We also prove a new bound on the LRC, which generally holds for hypothesis classes using strongly convex regularizers. This bound readily facilitates the bounding of the LRC for a range of such regularizers (not only for MTL, but also for the standard i.i.d. setting), as we demonstrate for classes induced by graph-based, Schatten- and group-norm regularizers. Moreover, we prove matching lower bounds showing that, aside from constants, the LRC-based bounds are tight for the considered applications.

Our derived bounds reflect that one can trade off a slow convergence speed w.r.t. T for an improved convergence rate w.r.t. n . The latter one ranges, in the worst case, from the typical GRC-based bounds $O(1/\sqrt{n})$, all the way up to the fastest rate of order $O(1/n)$ by allowing the bound to depend less on T . This trade-off is perhaps best exemplified in the case of Schatten norms, for which the two rates (exponents) always sum up to -1 . Nevertheless, the premium in question becomes less relevant to MTL, in which T is typically considered as fixed.

1.2 Organization of the paper

The paper is organized as follows: Sect. 2 lays the foundations for our analysis by considering a Talagrand-type concentration inequality suitable for deriving our bounds. Next, in Sect. 3, after suitably defining LRCs for MTL hypothesis spaces, we provide our LRC-based excess MTL risk bounds. Based on these bounds, we follow up this section with a local analysis of linear MTL frameworks, in which task-relatedness is presumed and enforced by imposing a strongly-convex norm constraint. In more detail, leveraging off the Fenchel-Young inequality, Sect. 4 presents a generic upper bound for the relevant LRC, which is subsequently specialized to the case of group norm, Schatten norm and graph-regularized linear MTL. After illustrating the tightness of the upper bounds, Sect. 5 supplies the corresponding excess risk bounds. The paper concludes with Sect. 6, which compares side by side the GRC- and LRC-based excess risk bounds for the aforementioned hypothesis spaces, as well as two additional related cases.

1.3 Previous Related Works

Earlier works that investigate MTL generalization guarantees employing Rademacher averages include [38], which considers linear MTL frameworks for binary classification. In these frameworks, all tasks are pre-processed by a common bounded linear operator and operator norm constraints are used to control the complexity of the associated hypothesis spaces. The GRC-based error bounds derived are of order $O(1/\sqrt{n})$ and non-vanishing w.r.t. T in the distribution-dependent case and of order $O(1/\sqrt{nT})$ in the data-dependent case. Another study, [39], provides bounds for the empirical and expected Rademacher complexities of

linear transformation classes. Based on Hölder’s inequality, GRC-based risk bounds of order $O(1/\sqrt{n})$ and non-vanishing w.r.t. T are established for MTL hypothesis spaces with graph-based and L_{S_q} -Schatten norm regularizers, where $q \in \{2\} \cup [4, \infty]$.

The subject of MTL generalization guarantees experienced renewed attention in more years. In [21], the authors take advantage of the strongly-convex nature of certain matrix-norm regularizers to easily obtain generalization bounds for a variety of machine learning problems. Part of their work is devoted to the realm of online and off-line MTL. In the latter case, which pertains to the focus of our work, the paper provides a distribution-dependent GRC-based excess risk bound of order $O(1/\sqrt{nT})$. Moreover, [41] presents a global Rademacher complexity analysis leading to both data and distribution-dependent excess risk bounds of order $O(\sqrt{\log(n)/n})$ and non-vanishing w.r.t. T for a trace norm regularized MTL model. Also, [40] examines the bounding of (global) Gaussian complexities of function classes that result from considering composite maps, as it is typical in MTL among other settings. An application of the paper’s results yields data-dependent MTL risk bounds of order $O(1/\sqrt{n})$ and non-vanishing w.r.t. T . More recently, [42] presents excess risk bounds for both MTL and Learning-To-Learn (LTL) settings and reveals conditions, under which MTL is more beneficial over learning tasks independently. The accompanying bounds are of order $O(1/\sqrt{nT})$ and, compared to the results in [39, 41], it enjoys the advantage of vanishing as $T \rightarrow +\infty$.

Finally, due to being domains related to MTL, but, at the same time, less connected to the focus of this paper, we only mention in passing a few works that pertain to generalization guarantees in the realm of life-long learning and domain adaptation. Generalization performance analysis in life-long learning has been investigated in [51, 11, 10, 46] and [45]. Also, in the context of domain adaptation, similar considerations are examined in [34, 36, 35, 18, 54, 37] and [19].

1.4 Basic Assumptions & Notation

Consider T supervised learning tasks sampled from the same input-output space $\mathcal{X} \times \mathcal{Y}$. Each task t is presented by an independent random variable (X_t, Y_t) governed by a probability distribution μ_t . Also, the *i.i.d.* samples related to each task t are described by the sequence $(X_t^i, Y_t^i)_{i=1}^n$, which is distributed according to μ_t .

In what follows, we use the following notational conventions: vectors and matrices are depicted in bold face. The superscript T , when applied to a vector/matrix, denotes the transpose of that quantity. We define $\mathbb{N}_S := \{1, \dots, S\}$. For any random variables X, Y and functions f we use $\mathbb{E}f(X, Y)$ and $\mathbb{E}_X f(X, Y)$ to denote the expectation w.r.t. all the involved random variables and the conditional expectation w.r.t. the random variable X . For any vector-valued function $\mathbf{f} = (f_1, \dots, f_T)$, we introduce the following two notations:

$$P\mathbf{f} := \frac{1}{T} \sum_{t=1}^T P f_t = \frac{1}{T} \sum_{t=1}^T \mathbb{E}(f(X_t)), \quad P_n \mathbf{f} := \frac{1}{T} \sum_{t=1}^T P_n f_t = \frac{1}{T} \sum_{t=1}^T \frac{1}{n} \sum_{i=1}^n f(X_t^i).$$

For any loss function ℓ and any $\mathbf{f} = (f_1, \dots, f_T)$ we define ℓ_{f_t} the function defined by $\ell_{f_t}((X_t, Y_t)_{t=1}^T) = \ell(f_t(X_t), Y_t)$.

2 Talagrand-Type Inequality for Multi-Task Learning

The derivation of our LRC-based error bounds for MTL is founded on the following modified Talagrand’s concentration inequality [13, 50] adapted to the context of MTL, showing that the uniform deviation between the true and empirical means in a vector-valued function class \mathcal{F} can be dominated by the associated *multi-task Rademacher complexity* plus a term involving the variance of functions in \mathcal{F} . We defer the proof in Appendix.

Theorem 1 (TALAGRAN-TYPE INEQUALITY FOR MTL). *Let $\mathcal{F} = \{\mathbf{f} := (f_1, \dots, f_T)\}$ be a class of vector-valued functions satisfying $\sup_{t,x} |f_t(x)| \leq b$. Let $X := (X_t^i)_{(t,i)=(1,1)}^{(T,N_t)}$ be a vector of $\sum_{t=1}^T N_t$ independent random variables where $X_t^1, \dots, X_t^{N_t}, \forall t$ are identically distributed. Let $\{\sigma_t^i\}_{t,i}$ be a sequence of independent Rademacher variables. If $\frac{1}{T} \sum_{t=1}^T \sup_{\mathbf{f} \in \mathcal{F}} \mathbb{E} [f_t(X_t^1)]^2 \leq r$, then, for every $x > 0$, with probability at least*

$$1 - e^{-x},$$

$$\sup_{\mathbf{f} \in \mathcal{F}} (P\mathbf{f} - P_n\mathbf{f}) \leq 4\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{2xr}{nT}} + \frac{8bx}{3nT}, \quad (1)$$

where $n := \min_{t \in \mathbb{N}_T} N_t$, and the multi-task Rademacher complexity of function class \mathcal{F} is defined as

$$\mathfrak{R}(\mathcal{F}) := \mathbb{E}_{X, \sigma} \left\{ \sup_{\mathbf{f}=(f_1, \dots, f_T) \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^T \frac{1}{N_t} \sum_{i=1}^{N_t} \sigma_t^i f_t(X_t^i) \right\}.$$

Note that the same bound also holds for $\sup_{\mathbf{f} \in \mathcal{F}} (P_n\mathbf{f} - P\mathbf{f})$.

In Theorem 1, the data from different tasks assumed to be mutually independent, which is typical in the MTL setting [38]. To present the results in a clear way we always assume in the following that the available data for each task is the same, namely n .

3 Excess MTL Risk Bounds based on Local Rademacher Complexities

Theorem 1 motivates us to extend the classical LRC $\mathfrak{R}(\mathcal{F}, r)$ for a scalar-valued function class \mathcal{F} : $\mathfrak{R}(\mathcal{F}, r) := \mathbb{E}_{X, \sigma} [\sup_{\mathbf{f} \in \mathcal{F}, V(\mathbf{f}) \leq r} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i)]$ to the Multi-Task Local Rademacher Complexity (MT-LRC) $\mathfrak{R}(\mathcal{F}, r)$ and its empirical counterpart $\hat{\mathfrak{R}}(\mathcal{F}, r)$ for a vector-valued function class \mathcal{F} as follows:

$$\mathfrak{R}(\mathcal{F}, r) := \mathbb{E} \left[\sup_{\substack{\mathbf{f}=(f_1, \dots, f_T) \in \mathcal{F} \\ V(\mathbf{f}) \leq r}} \frac{1}{nT} \sum_{\substack{t \in \mathbb{N}_T \\ i \in \mathbb{N}_n}} \sigma_t^i f_t(X_t^i) \right], \quad \hat{\mathfrak{R}}(\mathcal{F}, r) := \mathbb{E}_{\sigma} \left[\sup_{\substack{\mathbf{f}=(f_1, \dots, f_T) \in \mathcal{F} \\ V_n(\mathbf{f}) \leq r}} \frac{1}{nT} \sum_{\substack{t \in \mathbb{N}_T \\ i \in \mathbb{N}_n}} \sigma_t^i f_t(X_t^i) \right], \quad (2)$$

where $V(\mathbf{f})$ and $V_n(\mathbf{f})$ are upper bounds on the variance and conditional variances of the functions in \mathcal{F} , respectively. This paper takes the choice $V(\mathbf{f}) = P\mathbf{f}^2$ and $V_n(\mathbf{f}) = P_n\mathbf{f}^2$.

Analogous to single task learning, the challenge in applying MT-LRC (2) to refine the existing learning rates is to find an optimal radius trading-off the size of the set $\{\mathbf{f} \in \mathcal{F} : V(\mathbf{f}) \leq r\}$ and its complexity, which, as we show below, reduces to the calculation of the fixed-point of a sub-root function. We call a function ψ *sub-root* if it is non-decreasing, non-negative and $r \mapsto \psi(r)/\sqrt{r}$ is non-increasing for $r \geq 0$. We call the unique solution of the equation $\psi(r) = r$ the *fixed point* of ψ . We suppose that the loss function ℓ and the hypothesis space \mathcal{F} satisfy the following conditions:

Assumptions 1.

1. There is a function $\mathbf{f}^* = (f_1^*, \dots, f_T^*) \in \mathcal{F}$ satisfying $P\ell_{\mathbf{f}^*} = \inf_{\mathbf{f} \in \mathcal{F}} P\ell_{\mathbf{f}}$.
2. There is constant $B' \geq 1$, such that for every $f_t \in \mathcal{F}$ we have $P(f_t - f_t^*)^2 \leq B'P(\ell_{f_t} - \ell_{f_t^*})$.
3. There is a constant L , such that the loss function ℓ is L -Lipschitz in its first argument.

We now present the main result of this section showing that the excess error of MTL can be bounded by the fixed-point of a sub-root function dominating the MT-LRC.

Theorem 2 (Distribution-dependent excess risk bound for MTL). *Let $\mathcal{F} = \{\mathbf{f} := (f_1, \dots, f_T)\}$ be a class of vector-valued functions satisfying $\sup_{t,x} |f_t(x)| \leq b$. Let $X := (X_t^i, Y_t^i)_{(t,i)=(1,1)}^{(T,n)}$ be a vector of nT independent random variables where $(X_t^1, Y_t^1), \dots, (X_t^n, Y_t^n), \forall t$ are identically distributed. Suppose that Assumptions 1 holds. Let ψ be a sub-root function with the fixed point r^* such that $BL\mathfrak{R}(\mathcal{F}^*, r) \leq \psi(r), \forall r \geq r^*$, where $\mathfrak{R}(\mathcal{F}^*, r)$ is the LRC of the functions class \mathcal{F}^* :*

$$\mathfrak{R}(\mathcal{F}^*, r) := \mathbb{E}_{X, \sigma} \left[\sup_{\mathbf{f} \in \mathcal{F}, L^2 P(\mathbf{f} - \mathbf{f}^*)^2 \leq r} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \right]. \quad (3)$$

Then, we have the following bounds in terms of the fixed point r^* of $\psi(r)$:

1. For any function class \mathcal{F} , $K > 1$ and $x > 0$, with probability at least $1 - e^{-x}$,

$$P(\ell_{\mathbf{f}} - \ell_{\mathbf{f}^*}) \leq \frac{K}{K-1} P_n(\ell_{\mathbf{f}} - \ell_{\mathbf{f}^*}) + \frac{500K}{B} r^* + \frac{(6Lb + 10BK)x}{nT}.$$

2. For any convex function class \mathcal{F} , $K > 1$ and $x > 0$, with probability at least $1 - e^{-x}$,

$$P(\ell_{\mathbf{f}} - \ell_{\mathbf{f}^*}) \leq \frac{K}{K-1} P_n(\ell_{\mathbf{f}} - \ell_{\mathbf{f}^*}) + \frac{32K}{B} r^* + \frac{(3Lb + 4BK)x}{nT}.$$

Proof. Introduce the following class of excess loss functions:

$$\mathcal{H}_{\mathcal{F}}^* := \left\{ h_{\mathbf{f}} : (X_t, Y_t)_{t=1}^T \mapsto (\ell(f_t(X_t), Y_t) - \ell(f_t^*(X_t), Y_t))_{t=1}^T, \mathbf{f} = (f_1, \dots, f_T) \in \mathcal{F} \right\}. \quad (4)$$

From Assumptions 1, it can be seen that for any function $\mathbf{h} \in \mathcal{H}_{\mathcal{F}}^*$, $P(\ell_{\mathbf{f}} - \ell_{\mathbf{f}^*})^2 \leq L^2 P(\mathbf{f} - \mathbf{f}^*)^2 \leq BP(\ell_{\mathbf{f}} - \ell_{\mathbf{f}^*})$, where $B = B'L^2$. This implies

$$V(\mathbf{h}_{\mathbf{f}}) = Ph_{\mathbf{f}}^2 \leq L^2 P(\mathbf{f} - \mathbf{f}^*)^2 \leq BP(\mathbf{f} - \mathbf{f}^*).$$

Also, using Talagrand's Lemma [29], one can verify

$$\begin{aligned} B\mathfrak{R}(\mathcal{H}_{\mathcal{F}}^*, r) &= B\mathbb{E}_{X, \sigma} \left[\sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ V(\mathbf{h}_{\mathbf{f}}) \leq r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i h_t(X_t^i, Y_t^i) \right] \\ &= B\mathbb{E}_{X, \sigma} \left[\sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ V(\mathbf{h}_{\mathbf{f}}) \leq r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i \ell_{f_t}(X_t^i, Y_t^i) \right] \\ &\leq BL\mathfrak{R}(\mathcal{F}^*, r) \leq \psi(r). \end{aligned}$$

Applying Theorem A.2 (which is the extension of Theorem 3.3 of [7] to MTL function classes) to the function class $\mathcal{H}_{\mathcal{F}}^*$ completes the proof. \square

The following excess-risk bound is immediate by noting that $P_n(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq 0$.

Corollary 3. Let $\hat{\mathbf{f}}$ be any element of \mathcal{F} satisfying $P_n \ell_{\hat{\mathbf{f}}} = \inf_{\mathbf{f} \in \mathcal{F}} P_n \ell_{\mathbf{f}}$. Assume that the conditions of Theorem 2 hold. Then for any $x > 0$ and $r > \psi(r)$, with probability at least $1 - e^{-x}$,

$$P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq \frac{500K}{B} r^* + \frac{(6Lb + 10BK)x}{nT}. \quad (5)$$

The next theorem, analogous to Theorem 5.4 in [7], presents a data-dependent version of (5) replacing the Rademacher complexity in Corollary 3 with its empirical counterpart.

Theorem 4 (Data-dependent excess risk bound for MTL). Let $\hat{\mathbf{f}}$ be any element of \mathcal{F} satisfying $P_n \ell_{\hat{\mathbf{f}}} = \inf_{\mathbf{f} \in \mathcal{F}} P_n \ell_{\mathbf{f}}$. Assume that the condition of Theorem 2 hold. Define

$$\hat{\psi}_n(r) = c_1 \hat{\mathfrak{R}}(\mathcal{F}^*, c_3 r) + \frac{c_2 x}{nT}, \quad \hat{\mathfrak{R}}(\mathcal{F}^*, c_3 r) := \mathbb{E}_{\sigma} \left[\sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ L^2 P_n(\mathbf{f} - \hat{\mathbf{f}})^2 \leq c_3 r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \right],$$

where $c_1 = 2L \max(B, 16Lb)$, $c_2 = 8L^2 b^2 + c_1$ and $c_3 = 4 + 128K + 4B(3Lb + 4BK)/c_2$. Then for any $K > 1$ and $x > 0$, with probability at least $1 - 4e^{-x}$, we have

$$P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq \frac{500K}{B} \hat{r}^* + \frac{(6Lb + 10BK)x}{nT},$$

where \hat{r}^* is the fixed point of the sub-root function $\hat{\psi}_n(r)$.

Proof. The proof of this Theorem repeats the same basic steps utilized in Theorem 5.4 in [7] and, therefore, can be found in the Appendix. \square

4 Local Rademacher Complexity Bounds for MTL models with Strongly Convex Regularizers

This section presents very general MT-LRC bounds, based on the distribution-dependent excess risks established in Theorem 2, for hypothesis spaces defined by strongly convex regularizers, which allows us to immediately derive, as specific application cases, LRC bounds for group-norm, Schatten-norm, and graph-regularized MTL models. It should be mentioned that similar data-dependent MT-LRC bounds are also available by a similar deduction process.

4.1 Preliminaries

We consider linear MTL models where we associate to each task-wise function f_t a weight $\mathbf{w}_t \in \mathcal{H}$ by $f_t(X) = \langle \mathbf{w}_t, \phi(X) \rangle$. Here ϕ is a feature map associated to a Mercer kernel k satisfying $k(X, \tilde{X}) = \langle \phi(X), \phi(\tilde{X}) \rangle, \forall X, \tilde{X} \in \mathcal{X}$ and \mathbf{w}_t belongs to the *reproducing kernel Hilbert space* \mathcal{H}_K induced by k with inner product $\langle \cdot, \cdot \rangle$. We assume that the multi-task model $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_T) \in \mathcal{H} \times \dots \times \mathcal{H}$ is learned by a regularization scheme:

$$\min_{\mathbf{W}} R(\mathbf{W}) + C \sum_{t=1}^T \sum_{i=1}^n \ell(\langle \mathbf{w}_t, \phi(X_t^i) \rangle_{\mathcal{H}}, Y_t^i), \quad (6)$$

where the regularizer $R(\cdot)$ is used to enforce a priori information to avoid over-fitting. This regularization scheme amounts to performing *Empirical Risk Minimization (ERM)* in the hypothesis space

$$\mathcal{F} := \{X \mapsto [\langle \mathbf{w}_1, \phi(X_1) \rangle, \dots, \langle \mathbf{w}_T, \phi(X_T) \rangle]^T : R(\mathbf{W}) \leq R\}, \quad (7)$$

where for generality we consider regularizers of the form $R(\mathbf{W}) = \|\mathbf{D}^{1/2} \mathbf{W}\|$ with a positive operator \mathbf{D} defined in \mathcal{H} . The hypothesis space associated to group and Schatten norms can be recovered by taking $\mathbf{D} = \mathbf{I}$ and appropriate norms. Furthermore, the graph-regularized MTL can be specialized by taking $\mathbf{D} = \mathbf{L} + \eta \mathbf{I}$ with \mathbf{L} being a graph-Laplacian and η being a regularization constant. Our general discussion shows that all these MTL models can be covered in a framework with the notion of strong convexity.

4.2 General Bound on the LRC

Now, we can provide the main result of this section which gives a LRC bound for any general MTL hypothesis space of the form (7).

Theorem 5 (Distribution-dependent MT-LRC bounds by strong convexity). *Suppose that $R(\mathbf{W})$ in (6) is μ -strongly convex with $R^*(\mathbf{0}) = 0$ and $\|k\|_{\infty} \leq \beta \leq \infty$. Let X_t^1, \dots, X_t^n be an i.i.d. sample drawn from P_t . Also, assume that for each task t , the eigenvalue-eigenvector decomposition of the Hilbert-Schmidt covariance operator is given by $J_t = \mathbb{E}(\phi(X_t) \otimes \phi(X_t)) = \sum_{j=1}^{\infty} \lambda_t^j \mathbf{u}_t^j \otimes \mathbf{u}_t^j$, where $(\mathbf{u}_t^j)_{j=1}^{\infty}$ forms an orthonormal basis of \mathcal{H} , and $(\lambda_t^j)_{j=1}^{\infty}$ are the corresponding eigenvalues, for the task t , arranged in non-increasing order. Then for every positive operator \mathbf{D} on \mathbb{R}^T , any $r > 0$ and any non-negative integers h_1, \dots, h_T :*

$$\mathfrak{R}(\mathcal{F}, r) \leq \min_{0 \leq h_t \leq \infty} \left\{ \sqrt{\frac{r \sum_{t=1}^T h_t}{nT}} + \frac{1}{T} \sqrt{\frac{2R}{\mu} \mathbb{E}_{X, \sigma} \|\mathbf{D}^{-1/2} \mathbf{V}\|_*^2} \right\}, \quad (8)$$

where $\mathbf{V} = \left(\sum_{j > h_t} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right)_{t=1}^T$.

Proof. Note that with the help of LRC definition, we have for any function class \mathcal{F} ,

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}, r) &= \frac{1}{nT} \mathbb{E}_{X, \sigma} \left\{ \sup_{\substack{\mathbf{f}=(f_1, \dots, f_T) \in \mathcal{F}, \\ P\mathbf{f}^2 \leq r}} \sum_{i=1}^n \left\langle (\mathbf{w}_t)_{t=1}^T, (\sigma_t^i \phi(X_t^i))_{t=1}^T \right\rangle \right\} \\
&= \frac{1}{T} \mathbb{E}_{X, \sigma} \left\{ \sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ P\mathbf{f}^2 \leq r}} \left\langle (\mathbf{w}_t)_{t=1}^T, \left(\sum_{j=1}^{\infty} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right)_{t=1}^T \right\rangle \right\} \\
&\leq \frac{1}{T} \mathbb{E}_{X, \sigma} \left\{ \underbrace{\sup_{P\mathbf{f}^2 \leq r} \left\langle \left(\sum_{j=1}^{h_t} \sqrt{\lambda_t^j} \left\langle \mathbf{w}_t, \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right)_{t=1}^T, \left(\sum_{j=1}^{h_t} \sqrt{\lambda_t^j}^{-1} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right)_{t=1}^T \right\rangle}_{A_1} \right\}
\end{aligned} \tag{9}$$

$$+ \frac{1}{T} \mathbb{E}_{X, \sigma} \left\{ \underbrace{\sup_{\mathbf{f} \in \mathcal{F}} \left\langle (\mathbf{w}_t)_{t=1}^T, \left(\sum_{j>h_t} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right)_{t=1}^T \right\rangle}_{A_2} \right\}. \tag{10}$$

Step 1. Controlling A_1 : Applying Cauchy-Schwartz (C.S.) and Hölder inequalities on A_1 yields the following

$$\begin{aligned}
A_1 &\leq \frac{1}{T} \mathbb{E}_{X, \sigma} \left\{ \sup_{P\mathbf{f}^2 \leq r} \left[\left(\sum_{t=1}^T \left\| \sum_{j=1}^{h_t} \sqrt{\lambda_t^j} \left\langle \mathbf{w}_t, \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right\|^2 \right)^{\frac{1}{2}} \right. \right. \\
&\quad \left. \left. \left(\sum_{t=1}^T \left\| \sum_{j=1}^{h_t} \sqrt{\lambda_t^j}^{-1} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right\|^2 \right)^{\frac{1}{2}} \right] \right\} \\
&= \frac{1}{T} \mathbb{E}_{X, \sigma} \left\{ \sup_{P\mathbf{f}^2 \leq r} \left[\left(\sum_{t=1}^T \sum_{j=1}^{h_t} \lambda_t^j \left\langle \mathbf{w}_t, \mathbf{u}_t^j \right\rangle^2 \right)^{\frac{1}{2}} \right. \right. \\
&\quad \left. \left. \left(\sum_{t=1}^T \sum_{j=1}^{h_t} \lambda_t^{j-1} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \mathbf{u}_t^j \right\rangle^2 \right)^{\frac{1}{2}} \right] \right\}.
\end{aligned}$$

With the help of Jensen's inequality and regarding the fact that $\mathbb{E}_{X, \sigma} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \mathbf{u}_t^j \right\rangle^2 = \frac{\lambda_t^j}{n}$ and $P\mathbf{f}^2 \leq r$ implies $\frac{1}{T} \sum_{t=1}^T \sum_{j=1}^{\infty} \lambda_t^j \left\langle \mathbf{w}_t, \mathbf{u}_t^j \right\rangle^2 \leq r$ (see Lemma 2 in the Appendix for the proof), we can further bound A_1 as

$$A_1 \leq \sqrt{\frac{r \sum_{t=1}^T h_t}{nT}}. \tag{11}$$

Step 2. Controlling A_2 : We use strong convexity assumption on the regularizer in order to further bound the second term A_2 .

Let $\lambda > 0$. Applying (A.27) with $\mathbf{w} = \mathbf{D}^{1/2} \mathbf{W}$ and $\mathbf{v} = \lambda \mathbf{D}^{-1/2} \mathbf{V}$ gives

$$\left\langle \mathbf{D}^{1/2} \mathbf{W}, \lambda \mathbf{D}^{-1/2} \mathbf{V} \right\rangle \leq R(\mathbf{W}) + \left\langle \nabla R^*(\mathbf{0}), \lambda \mathbf{D}^{-1/2} \mathbf{V} \right\rangle + \frac{\lambda^2}{2\mu} \left\| \mathbf{D}^{-1/2} \mathbf{V} \right\|_*^2.$$

Taking expectation on both sides and optimizing λ gives

$$\begin{aligned} A_2 &= \frac{1}{T} \mathbb{E}_{X, \sigma} \left\{ \sup_{\mathbf{f} \in \mathcal{F}} \left\langle \mathbf{D}^{1/2} \mathbf{W}, \mathbf{D}^{-1/2} \mathbf{V} \right\rangle \right\} \leq \min_{0 < \lambda < \infty} \left\{ \frac{R}{\lambda T} + \frac{\lambda}{2\mu T} \mathbb{E}_{X, \sigma} \left\| \mathbf{D}^{-1/2} \mathbf{V} \right\|_*^2 \right\} \\ &= \frac{1}{T} \sqrt{\frac{2R}{\mu} \mathbb{E}_{X, \sigma} \left\| \mathbf{D}^{-1/2} \mathbf{V} \right\|_*^2}. \end{aligned} \quad (12)$$

Combining (12) with (11) completes the proof. \square

In the following we demonstrate the power of Theorem 5 by applying it to study LRC bounds for popular MTL models, including group norm, Schatten norm and graph regularized MTL models extensively studied in the literature of MTL [39, 20, 6, 4, 31, 3].

4.3 Group Norm Regularized MTL

We first consider the group norm regularized MTL capturing the inter-task relationships by the group norm $\|\mathbf{W}\|_{2,q} := (\sum_{t=1}^T \|\mathbf{w}_t\|_2^q)^{1/q}$ [20, 4, 32, 49], for which the associated hypothesis space takes the form

$$\mathcal{F}_q := \left\{ X \mapsto [\langle \mathbf{w}_1, \phi(X_1) \rangle, \dots, \langle \mathbf{w}_T, \phi(X_T) \rangle]^T : \|\mathbf{W}\|_{2,q} \leq R_{max} \right\}. \quad (13)$$

Corollary 6. *If $1 \leq q \leq 2$ in (13), the LRC of function class \mathcal{F}_q can be bounded as*

$$\mathfrak{R}(\mathcal{F}_q, r) \leq \sqrt{\frac{4}{nT} \left\| \left(\sum_{j=1}^{\infty} \min \left(rT^{1-\frac{2}{q^*}}, \frac{2eq^{*3}R_{max}}{T} \lambda_t^j \right) \right)_{t=1}^T \right\|_{\frac{q^*}{2}}^T} + \frac{\sqrt{2\beta e R_{max} q^{*\frac{3}{2}} T^{\frac{1}{q^*}}}}{nT}. \quad (14)$$

Proof. The proof follows by applying Khintchine (A.28) and Rosenthal (A.29) inequalities to further bound the expectation term in (8) which gives,

$$A_2(\mathcal{F}_q) \leq \sqrt{\frac{2eq^{*2}R_{max}}{nT^2\mu} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}^T} + \frac{\sqrt{2\beta e R_{max} q^{*} T^{\frac{1}{q^*}}}}{nT\sqrt{\mu}}. \quad (15)$$

Now, combining (11) and (15) provides the bound on $\mathfrak{R}(\mathcal{F}_q, r)$ as

$$\mathfrak{R}(\mathcal{F}_q, r) \leq \sqrt{\frac{r \sum_{t=1}^T h_t}{nT}} + \sqrt{\frac{2eq^{*2}R_{max}}{nT^2\mu} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}^T} + \frac{\sqrt{2\beta e R_{max} q^{*} T^{\frac{1}{q^*}}}}{nT\sqrt{\mu}}, \quad (16)$$

Then using the following inequalities according which for any non-negative numbers $\alpha_1, \alpha_2 \in \mathbb{R}^+$ and any non-negative vectors $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^T$ with $0 \leq q \leq p \leq \infty$ any $s \geq 1$,

$$(\star) \quad \sqrt{\alpha_1} + \sqrt{\alpha_2} \leq \sqrt{2(\alpha_1 + \alpha_2)} \quad (17)$$

$$(\star\star) \quad l_p - t_0 - l_q : \quad \|\mathbf{a}_1\|_q = \langle \mathbf{1}, \mathbf{a}_1 \rangle^{\frac{1}{q}} \stackrel{\text{H\"older}}{\leq} \left(\|\mathbf{1}\|_{(p/q)^*} \|\mathbf{a}_1^q\|_{(p/q)} \right)^{\frac{1}{q}} = T^{\frac{1}{q} - \frac{1}{p}} \|\mathbf{a}_1\|_p \quad (18)$$

$$(\star\star\star) \quad \|\mathbf{a}_1\|_s + \|\mathbf{a}_2\|_s \leq 2^{1-\frac{1}{s}} \|\mathbf{a}_1 + \mathbf{a}_2\|_s \leq 2 \|\mathbf{a}_1 + \mathbf{a}_2\|_s, \quad (19)$$

we can obtain the desired result. See the Appendix for the detailed proof. \square

Remark 7. Since the LRC bound above is not monotonic in q it is more reasonable to state the above bound in terms of $q \leq \kappa$, as taking $\kappa = q$ is not always the optimal choice. Trivially for the group norm regularizer with any $\kappa \geq q$, it holds that $\|\mathbf{W}\|_{2,\kappa} \leq \|\mathbf{W}\|_{2,q}$ and therefore $\mathfrak{R}(\mathcal{F}_q, r) \leq \mathfrak{R}(\mathcal{F}_\kappa, r)$. Thus we have the following bound on $\mathfrak{R}(\mathcal{F}_q, r)$ for any $\kappa \in [q, 2]$,

$$\mathfrak{R}(\mathcal{F}_q, r) \leq \sqrt{\frac{4}{nT} \left\| \left(\sum_{j=1}^{\infty} \min \left(rT^{1-\frac{2}{\kappa^*}}, \frac{2e\kappa^{*3}R_{max}}{T} \lambda_t^j \right) \right)_{t=1}^T \right\|_{\frac{\kappa^*}{2}}} + \frac{\sqrt{2\beta e R_{max}} \kappa^{*\frac{3}{2}} T^{\frac{1}{\kappa^*}}}{nT}.$$

Remark 8 (Sparsity-inducing group-norm). Among $p \geq 1$, a particular group-norm of independent interest is the sparsity-inducing group-norm achieved by $q = 1$ [6, 4], for which we can take $\kappa^* = \log T$ to get that

$$\begin{aligned} \mathfrak{R}(\mathcal{F}_1, r) &\leq \sqrt{\frac{4}{nT} \left\| \left(\sum_{j=1}^{\infty} \min \left(rT^{1-\frac{2}{\kappa^*}}, \frac{2e\kappa^{*3}R_{max}}{T} \lambda_t^j \right) \right)_{t=1}^T \right\|_{\frac{\kappa^*}{2}}} + \frac{\sqrt{2\beta e R_{max}} \kappa^{*\frac{3}{2}} T^{\frac{1}{\kappa^*}}}{nT} \\ &\stackrel{(l_{\frac{\kappa^*}{2}} - to - l_{\infty})}{\leq} \sqrt{\frac{4}{nT} \left\| \left(\sum_{j=1}^{\infty} \min \left(rT, \frac{2e^3(\log T)^3}{T} R_{max} \lambda_t^j \right) \right)_{t=1}^T \right\|_{\infty}} + \frac{\sqrt{2\beta R_{max}} e^{\frac{3}{2}} (\log T)^{\frac{3}{2}}}{nT}. \end{aligned}$$

To investigate the tightness of the bound in (14), we derive the lower bound which holds for the LRC of \mathcal{F}_q with any $q \geq 1$.

Theorem 9 (Lower bound). *The following lower bound holds for the local Rademacher complexity of \mathcal{F}_q in (14) with any $q \geq 1$. There is an absolute constant c so that $\forall t$, if $\lambda_t^1 \geq 1/(nR_{max}^2)$ then for all $r \geq \frac{1}{n}$ and $q \geq 1$,*

$$\mathfrak{R}(\mathcal{F}_{q,R,T}, r) \geq \sqrt{\frac{c}{nT^{1-\frac{2}{q^*}}} \sum_{j=1}^{\infty} \min \left(rT^{1-\frac{2}{q^*}}, \frac{R_{max}^2}{T} \lambda_1^j \right)}. \quad (20)$$

Proof. The proof can be found in the Appendix. \square

A comparison between the lower bound in (20) and the upper bound in (14) can be clearly illustrated by assuming identical eigenvalue tail sums $\sum_{j \geq \infty} \lambda_t^j$ for all tasks, for which the upper bound translates to

$$\mathfrak{R}(\mathcal{F}_{q,R,T}, r) \leq \sqrt{\frac{4}{nT^{1-\frac{2}{q^*}}} \sum_{j=1}^{\infty} \min \left(rT^{1-\frac{2}{q^*}}, \frac{2eq^{*3}R_{max}}{T} \lambda_t^j \right)} + \frac{\sqrt{2\beta e R_{max}} q^{*\frac{3}{2}} T^{\frac{1}{q^*}}}{nT},$$

matching the lower bound in (20) up to constants of factors. A similar comparison analysis can be performed for MTL models with Schatten norm and graph regularizers.

4.4 Schatten Norm Regularized MTL

[6] developed a spectral regularization framework for MTL where the Schatten p -norm $\|\mathbf{W}\|_{S_q} := [\text{tr}(\mathbf{W}^T \mathbf{W})^{\frac{q}{2}}]^{\frac{1}{q}}$ is studied as a concrete example, corresponding to perform ERM in the following hypothesis space:

$$\mathcal{F}_{S_q} := \{X \mapsto [\langle \mathbf{w}_1, \phi(X_1) \rangle, \dots, \langle \mathbf{w}_T, \phi(X_T) \rangle]^T : \|\mathbf{W}\|_{S_q} \leq R'_{max}\}. \quad (21)$$

Corollary 10. *For any $1 \leq q \leq 2$ in (21), the LRC of function class \mathcal{F}_{S_q} is bounded as*

$$\mathfrak{R}(\mathcal{F}_{S_q}, r) \leq \sqrt{\frac{4}{nT} \left\| \left(\sum_{j=1}^{\infty} \min \left(r, \frac{2q^*(q^*-1)R'_{max}}{T} \lambda_t^j \right) \right)_{t=1}^T \right\|_1}.$$

Proof. The proof is provided in the Appendix. \square

Remark 11 (Sparsity-inducing Schatten-norm (trace norm)). Trace-norm regularized MTL, corresponding to Schatten norm regularization with $q = 1$ [41, 48], imposes a low-rank structure on the spectrum of \mathbf{W} and can also be interpreted as low dimensional subspace learning [5, 28, 22]. Note that for any $q \geq 1$, it holds that $\mathfrak{R}(\mathcal{F}_{S_1}, r) \leq \mathfrak{R}(\mathcal{F}_{S_q}, r)$. Therefore, choosing the optimal $q^* = 2$ yields that

$$\mathfrak{R}(\mathcal{F}_{S_1}, r) \leq \sqrt{\frac{4}{nT} \left\| \left(\sum_{j=1}^{\infty} \min \left(r, \frac{4R'_{max}}{T} \lambda_t^j \right) \right)_{t=1}^T \right\|_1}.$$

4.5 Graph Regularized MTL

The idea underlying graph regularized MTL is to force the classifiers of related tasks close to each other by penalizing the squared distance $\|\mathbf{w}_t - \mathbf{w}_s\|^2$ with different weights ω_{ts} . We consider the following graph regularized MTL [39]

$$R(\mathbf{W}) = \frac{1}{2T} \sum_{t=1}^T \sum_{s=1}^T \omega_{ts} \|\mathbf{w}_t - \mathbf{w}_s\|^2 + \frac{\eta}{T} \sum_{t=1}^T \|\mathbf{w}_t\|^2 = \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T (\mathbf{L} + \eta \mathbf{I})_{ts} \langle \mathbf{w}_t, \mathbf{w}_s \rangle,$$

where \mathbf{L} is the graph-Laplacian associated to a matrix of edge-weights $(\omega_{ts})_{ts}$, \mathbf{I} is the identity in \mathbb{R}^T , and $\eta > 0$ is a regularization parameter. According to the identity $\frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T (\mathbf{L} + \eta \mathbf{I})_{ts} \langle \mathbf{w}_t, \mathbf{w}_s \rangle = (1/T) \|(\mathbf{L} + \eta \mathbf{I})^{1/2} \mathbf{W}\|_F^2$, the corresponding hypothesis space is:

$$\mathcal{F}_G := \left\{ X \mapsto [\langle \mathbf{w}_1, \phi(X_1) \rangle, \dots, \langle \mathbf{w}_T, \phi(X_T) \rangle]^T : \|\mathbf{D}^{1/2} \mathbf{W}\|_F \leq R''_{max} \right\}. \quad (22)$$

Corollary 12. For any positive definite matrix \mathbf{D} in (22), the LRC of \mathcal{F}_G is bounded by

$$\mathfrak{R}(\mathcal{F}_G, r) \leq \sqrt{\frac{4}{nT} \left\| \left(\sum_{j=1}^{\infty} \min \left(r, \frac{2\mathbf{D}_{tt}^{-1} R''_{max}}{T} \lambda_t^j \right) \right)_{t=1}^T \right\|_1}. \quad (23)$$

Proof. See the Appendix for the proof. \square

5 Excess Risk Bounds for MTL models with Strongly Convex Regularizers

In this section we will provide the distribution and data-dependent excess risk bounds for the hypothesis spaces considered earlier. Note that, due to the space limitation, the proofs of the results are provided only for the hypothesis space \mathcal{F}_q with $q \in [1, 2]$ in (13). However, for the group and L_{S_q} -Schatten norm ($q \in [1, 2]$) regularized MTL, the proofs can be obtained in a very similar way.

Theorem 13 (Distribution-dependent excess risk bound for a MTL problem with a strong convex $L_{2,q}$ group-norm regularizer). Assume the convex class \mathcal{F}_q in (13) has ranges in $[-b, b]$, and let the loss function ℓ in Problem (6) be such that Assumptions 1 are satisfied. Let $\hat{\mathbf{f}}$ be any element of \mathcal{F}_q with $1 \leq q \leq 2$ which satisfies $P_n \ell_{\hat{\mathbf{f}}} = \inf_{\mathbf{f} \in \mathcal{F}_q} P_n \ell_{\mathbf{f}}$. Assume moreover that k is a positive semi-definite kernel on \mathcal{X} such that $\|k\|_{\infty} \leq \beta \leq \infty$. Denote by r^* the fixed point of $2BL\mathfrak{R}(\mathcal{F}_q, \frac{r}{4L^2})$. Then, for any $K > 1$ and $x > 0$, with probability at least $1 - e^{-x}$, the excess loss of function class \mathcal{F}_q is bounded as

$$P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq \frac{32K}{B} r^* + \frac{(3Lb + 4BK)x}{nT}, \quad (24)$$

where for the fixed point r^* of the local Rademacher complexity $2BL\mathfrak{R}(\mathcal{F}_q, \frac{r}{4L^2})$, it holds

$$r^* \leq \min_{0 \leq h_t \leq \infty} \frac{B^2 \sum_{t=1}^T h_t}{Tn} + 4BL \sqrt{\frac{2eq^{*3} R_{max}}{nT^2} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q}{2}}} + \frac{4\sqrt{2\beta e R_{max}} q^{*\frac{3}{2}} T^{\frac{1}{q^*}}}{nT}, \quad (25)$$

where h_1, \dots, h_T are arbitrary non-negative integers.

Proof. First notice that \mathcal{F}_q is convex, thus it is star-shaped around any of its points. Hence according to Lemma 3.4 in [7], \mathcal{F}_q is a sub-root function. Moreover, because of the symmetry of σ_t^i and because \mathcal{F}_q is convex and symmetric, it can be shown that $\mathfrak{R}(\mathcal{F}_q^*, r) \leq 2\mathfrak{R}(\mathcal{F}_q, \frac{r}{4L^2})$, where $\mathfrak{R}(\mathcal{F}_q^*, r)$ is defined according to (3) for the class of functions \mathcal{F}_q . Therefore, it suffices to find the fixed point of $2BL\mathfrak{R}(\mathcal{F}_q, \frac{r}{4L^2})$ by solving $\phi(r) = r$. For this purpose, we will use (16) as a bound for $\mathfrak{R}(\mathcal{F}_q, r)$, and we solve $\sqrt{\alpha r} + \gamma = r$ which is equivalent to solving $r^2 - (\alpha + 2\gamma)r + \gamma^2 = 0$, where we define

$$\alpha = \frac{B^2 \sum_{t=1}^T h_t}{Tn}, \quad \text{and} \quad \gamma = 2BL \sqrt{\frac{2eq^{*3}R_{max}}{nT^2} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}} + \frac{2\sqrt{2\beta e R_{max}} BL q^{*\frac{3}{2}} T^{\frac{1}{q^*}}}{nT}. \quad (26)$$

It is not hard to verify that $r^* \leq \alpha + 2\gamma$. Substituting the definition of α and γ gives the result. \square

Regarding the fact that λ_t^j are decreasing with respect to j , we can assume $\exists d_t : \lambda_t^j \leq d_t j^{-\alpha_t}$ for some $\alpha_t > 1$. As examples, this assumption holds for finite rank kernels as well as convolution kernels. Thus, it can be shown

$$\sum_{j>h_t} \lambda_t^j \leq d_t \sum_{j>h_t} j^{-\alpha_t} \leq d_t \int_{h_t}^{\infty} x^{-\alpha_t} dx = d_t \left[\frac{1}{1-\alpha_t} x^{1-\alpha_t} \right]_{h_t}^{\infty} = -\frac{d_t}{1-\alpha_t} h_t^{1-\alpha_t}. \quad (27)$$

note that by $l_p - to - l_q$ conversion, we have

$$\frac{B^2 \sum_{t=1}^T h_t}{Tn} \leq B \sqrt{\frac{B^2 T \sum_{t=1}^T h_t^2}{n^2 T^2}} \leq B \sqrt{\frac{B^2 T^{2-\frac{2}{q^*}} \left\| (h_t^2)_{t=1}^T \right\|_{\frac{q^*}{2}}}{n^2 T^2}}.$$

Now, applying, (17) and (19), and inserting (27) into (25), it holds for a group norm regularized MTL with $1 \leq q \leq 2$,

$$r^* \leq \min_{0 \leq h_t \leq \infty} 2B \sqrt{\left\| \left(\frac{B^2 T^{2-\frac{2}{q^*}} h_t^2}{n^2 T^2} - \frac{32d_t e q^{*3} R_{max} L^2}{n T^2 (1-\alpha_t)} h_t^{1-\alpha_t} \right)_{t=1}^T \right\|_{\frac{q^*}{2}}} + \frac{4\sqrt{2\beta e R_{max}} BL q^{*\frac{3}{2}} T^{\frac{1}{q^*}}}{nT}. \quad (28)$$

Taking the derivative of the above bound with respect to h_t and setting it to zero yields the optimal h_t as

$$h_t = \left(16d_t e q^{*3} R_{max} B^{-2} L^2 T^{\frac{2}{q^*}-2} n \right)^{\frac{1}{1+\alpha_t}}.$$

Note that substituting the above for $\alpha := \min_{t \in \mathbb{N}_T} \alpha_t$ and $d = \max_{t \in \mathbb{N}_T} d_t$ into (28), we can upper-bound the fixed point of r^* as

$$r^* \leq \frac{8B^2}{n} \sqrt{e \frac{\alpha+1}{\alpha-1}} \left(d q^{*3} R_{max} B^{-2} L^2 T^{\frac{2}{q^*}-2} n \right)^{\frac{1}{1+\alpha}} + \frac{4\sqrt{2\beta e R_{max}} BL q^{*\frac{3}{2}} T^{\frac{1}{q^*}}}{nT}.$$

Also, the convergence rate of r^* can be determined as

$$r^* = O \left(\left(\frac{T^{2-\frac{2}{q^*}}}{q^{*3}} \right)^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} \right).$$

It can be seen that the convergence rate can be as slow as $O \left(\frac{q^{*3/2} T^{1/q^*}}{T\sqrt{n}} \right)$ (for small α where at least one $\alpha_t \approx 1$), and as fast as $O(n^{-1})$ (for large α where for all t , $\alpha_t \mapsto \infty$). Therefore, one can observe that the bound obtained for the fixed point together with Theorem 13 provides a bound for the excess risk which leads to the following theorem.

Remark 14 (Excess risk bounds for some strong convex matrix norm regularized MTL problems). Assume the convex class \mathcal{F}_q in (13) has ranges in $[-b, b]$, and let the loss function ℓ in Problem (6) be such that Assumptions 1 are satisfied. Assume moreover that k is a positive semidefinite kernel on \mathcal{X} such that $\|k\|_\infty \leq \beta \leq \infty$. Also, denote $\alpha := \min_{t \in \mathbb{N}_T} \alpha_t$ and $d = \max_{t \in \mathbb{N}_T} d_t$. Also,

- Group norm ($1 \leq q \leq 2$): If $\hat{\mathbf{f}}$ satisfies $P_n \ell_{\hat{\mathbf{f}}} = \inf_{\mathbf{f} \in \mathcal{F}_q} P_n \ell_{\mathbf{f}}$, and r^* is the fixed point of the local Rademacher complexity $2BL\mathfrak{R}(\mathcal{F}_q, \frac{r}{4L^2})$ with any $1 \leq q \leq 2$ in (13) and any $K > 1$, it holds with probability at least $1 - e^{-x}$,

$$P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq \min_{\kappa \in [q, 2]} 423K \sqrt{\frac{\alpha+1}{\alpha-1}} \left(d\kappa^{*3} R_{\max} L^2 \right)^{\frac{1}{1+\alpha}} B^{\frac{\alpha-1}{\alpha+1}} \left(T^{\frac{2}{\kappa^*}-2} \right)^{\frac{1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} \\ + \frac{299\sqrt{\beta R_{\max}} K L \kappa^{*\frac{3}{2}} T^{\frac{1}{\kappa^*}}}{nT} + \frac{(3Lb + 4BK)x}{nT}. \quad (29)$$

- Schatten-norm ($1 \leq q \leq 2$): If $\hat{\mathbf{f}}$ satisfies $P_n \ell_{\hat{\mathbf{f}}} = \inf_{\mathbf{f} \in \mathcal{F}_{S_q}} P_n \ell_{\mathbf{f}}$, and r^* is the fixed point of the local Rademacher complexity $2BL\mathfrak{R}(\mathcal{F}_{S_q}, \frac{r}{4L^2})$ with any $1 \leq q \leq 2$ in (21) and any $K > 1$, it holds with probability at least $1 - e^{-x}$,

$$P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq 256K \sqrt{\frac{\alpha+1}{\alpha-1}} \left(dq^*(q^*-1) R'_{\max} L^2 \right)^{\frac{1}{1+\alpha}} B^{\frac{\alpha-1}{\alpha+1}} T^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} \\ + \frac{(3Lb + 4BK)x}{nT}. \quad (30)$$

- Graph regularizer: If $\hat{\mathbf{f}}$ satisfies $P_n \ell_{\hat{\mathbf{f}}} = \inf_{\mathbf{f} \in \mathcal{F}_q} P_n \ell_{\mathbf{f}}$, and r^* is the fixed point of the local Rademacher complexity $2BL\mathfrak{R}(\mathcal{F}_G, \frac{r}{4L^2})$ with any positive operator \mathbf{D} in (22) and any $K > 1$, it holds with probability at least $1 - e^{-x}$,

$$P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq 256K \sqrt{\frac{\alpha+1}{\alpha-1}} \left(dR''_{\max} L^2 \mathbf{D}_{\max}^{-1} \right)^{\frac{1}{1+\alpha}} B^{\frac{\alpha-1}{\alpha+1}} T^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}} \\ + \frac{(3Lb + 4BK)x}{nT}. \quad (31)$$

where $\mathbf{D}_{\max}^{-1} := \max_{t \in \mathbb{N}_T} \mathbf{D}_{tt}^{-1}$.

Corollary 15 (Data-dependent excess risk bound for a MTL problem with a strong convex $L_{2,q}$ group-norm regularizer). Assume the convex class \mathcal{F}_q in (13) has ranges in $[-b, b]$, and let the loss function ℓ in Problem (6) be such that Assumptions 1 are satisfied. Let $\hat{\mathbf{f}}$ be any element of \mathcal{F}_q with $1 \leq q \leq 2$ which satisfies $P_n \ell_{\hat{\mathbf{f}}} = \inf_{\mathbf{f} \in \mathcal{F}_q} P_n \ell_{\mathbf{f}}$. Assume moreover that k is a positive semidefinite kernel on \mathcal{X} such that $\|k\|_\infty \leq \beta \leq \infty$. Let \mathbf{K}_t be the $n \times n$ kernel Gram matrix of task t with entries $(\mathbf{K}_t)_{ij} := k(X_t^i, X_t^j)$; denote $\hat{\lambda}_t^1, \dots, \hat{\lambda}_t^n$ its ordered eigenvalues. Let \hat{r}^* be the fixed point of

$$\hat{\psi}_n(r) = c_1 \hat{\mathfrak{R}}(\mathcal{F}_q^*, c_3 r) + \frac{c_2 x}{nT},$$

where $c_1 = 2L \max(B, 16Lb)$, $c_2 = 8L^2 b^2 + c_1$ and $c_3 = 4 + 128K + 4B(3Lb + 4BK)/c_2$, and

$$\hat{\mathfrak{R}}(\mathcal{F}_q^*, c_3 r) := \mathbb{E}_\sigma \left[\sup_{\substack{\mathbf{f} \in \mathcal{F}_q, \\ L^2 P_n(\mathbf{f} - \hat{\mathbf{f}})^2 \leq c_3 r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \middle| \{x_t^i\}_{t \in \mathbb{N}_T, i \in \mathbb{N}_n} \right]. \quad (32)$$

Then, for any $K > 1$ and $x > 0$, with probability at least $1 - 4e^{-x}$ the excess loss of function class \mathcal{F}_q is bounded as

$$P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq \frac{32K}{B} \hat{r}^* + \frac{(3Lb + 4BK)x}{nT}, \quad (33)$$

where for the fixed point \hat{r}^* of the empirical local Rademacher complexity $\hat{\psi}_n(r)$, it holds

$$\hat{r}^* \leq \frac{c_1^2 c_3 \sum_{t=1}^T h_t}{nTL^2} + 4 \sqrt{\frac{2c_1^2 q^{*2} R_{max}}{nT^2} \left\| \left(\sum_{j>h_t}^n \hat{\lambda}_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}} + \frac{2c_2 x}{nT},$$

where h_1, \dots, h_T are arbitrary non-negative integers, and $(\hat{\lambda}_t^j)_{j=1}^n$ are eigenvalues of the empirical Gram matrix \mathbf{K} obtained from kernel function k .

Proof. The proof of the result is provided in the Appendix. \square

6 Discussion

6.1 Global vs. Local Rademacher Complexity Bounds

This section is devoted to compare the excess risk bounds based on local Rademacher complexity to those of the global ones.

First, note that to obtain the GRC-based bounds, we apply Theorem 16 of [38], as we consider the same setting and assumptions for tasks' distributions as considered in this work. This theorem presents an MTL bound based on the notion of GRC.

Theorem 16 (MTL excess risk bound based on GRC; Theorem 16 of [38]). *Let \mathcal{F} be a class of vector-valued functions $\mathbf{f} = (f_1, \dots, f_T) : \mathcal{X} \mapsto \mathbb{R}^T$, that maps \mathcal{X} into $[-b, b]^T$, and let $X = (X_i^t)_{(i,t)=(1,1)}^{(n,T)}$ be a vector of independent random variables where for all fixed t , X_1^t, \dots, X_n^t are identically distributed according to P_t . Then for every $x > 0$, with probability at least $1 - e^{-x}$,*

$$\sup_{\mathbf{f} \in \mathcal{F}} (P\mathbf{f} - P_n\mathbf{f}) \leq 2\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{bx}{nT}}. \quad (34)$$

Proof. As it has been shown in [38], the proof of this theorem is based on using McDiarmid's inequality for Z defined in Theorem 1, and noticing that for the function class \mathcal{F} with values in $[-b, b]$, it holds that $|Z - Z_{s,j}| \leq 2b/nT$. \square

It can be observed that, in order to obtain the excess risk bound in the above theorem, one has to bound the GRC term $\mathfrak{R}(\mathcal{F})$ in (34). Therefore, we first upper-bound the GRC of different hypothesis spaces considered in the previous sections.

Theorem 17 (Distribution-dependent GRC bounds). *Assume that the conditions of Theorem 5 hold. Then, the following results hold for the GRC of the hypothesis spaces in (13), (21) and (22), respectively.*

- *Group-norm regularizer: For any $q \geq 1$ in (13), the GRC of the function class \mathcal{F}_q can be bounded as*

$$\forall \kappa > q : \quad \mathfrak{R}(\mathcal{F}_q) \leq \sqrt{\frac{2e\kappa^{*3} R_{max}}{nT^2} \left\| (\mathbf{tr}(J_t))_{t=1}^T \right\|_{\frac{\kappa^*}{2}}} + \frac{\sqrt{2\beta e R_{max} \kappa^{*\frac{3}{2}} T^{\frac{1}{\kappa^*}}}}{nT}.$$

- *Schatten-norm regularizer: For any $q \geq 1$ in (21), the GRC of the function class \mathcal{F}_{S_q} can be bounded as*

$$\mathfrak{R}(\mathcal{F}_{S_q}) \leq \sqrt{\frac{2R'_{max} q^*(q^* - 1)}{nT^2} \left\| (\mathbf{tr}(J_t))_{t=1}^T \right\|_1}. \quad (35)$$

- *Graph regularizer: For any positive operator \mathbf{D} in (22), the GRC of the function class \mathcal{F}_G can be bounded as*

$$\mathfrak{R}(\mathcal{F}_G) \leq \sqrt{\frac{2R''_{max}}{nT^2} \left\| (\mathbf{D}_{tt}^{-1} \mathbf{tr}(J_t))_{t=1}^T \right\|_1}. \quad (36)$$

Proof. The proof of the results can be found in the Appendix. \square

Notice that, assuming a unique bound for the traces of all task's kernels, the bound above is determined by $O\left(\frac{q^* \frac{3}{2} T^{\frac{1}{q^*}}}{T\sqrt{n}}\right)$. Also, taking $q^* = \log T$, we obtain the bound $\mathfrak{R}(\mathcal{F}_1)$ of order $\frac{(\log T)^{\frac{3}{2}}}{T\sqrt{n}}$. We can also remark that when the traces of the kernels are bounded, the bounds for the Schatten norm and graph regularizers are the order of $O\left(\frac{1}{\sqrt{nT}}\right)$.

Note that for the purpose of comparison, we concentrate only on the parameters R, n, T, q^* and α and assume all the other parameters are fixed and hidden in the big- O notation. Also, for the sake of simplicity, we assume that the eigenvalues of all tasks satisfy $\lambda_t^j \leq d_j^{-\alpha}$ (with $\alpha \geq 1$). Note that from Theorem 16, it follows that a bound on the global Rademacher complexity provides also a bound on the excess risk. This together with Theorem 17, gives the GRC-based excess risk bounds of the following forms

$$\begin{aligned} \text{Group norm: } \forall \kappa \in [q, 2], \quad & P(\ell_{\hat{f}} - \ell_{f^*}) = O\left((R_{max}\kappa^{*3})^{\frac{1}{2}} \left(T^{2-\frac{2}{\kappa^*}}\right)^{-\frac{1}{2}} n^{-\frac{1}{2}}\right). \\ \text{Schatten-norm: } \forall q \in [1, 2], \quad & P(\ell_{\hat{f}} - \ell_{f^*}) = O\left((R'_{max}q^*(q^* - 1))^{\frac{1}{2}} T^{-\frac{1}{2}} n^{-\frac{1}{2}}\right). \\ \text{Graph regularizer:} \quad & P(\ell_{\hat{f}} - \ell_{f^*}) = O\left((R''_{max})^{\frac{1}{2}} T^{-\frac{1}{2}} n^{-\frac{1}{2}}\right). \end{aligned} \quad (37)$$

which can be compared to their LRC-based counterparts as following

$$\begin{aligned} \text{Group norm: } \forall \kappa \in [q, 2], \quad & P(\ell_{\hat{f}} - \ell_{f^*}) = O\left((R_{max}\kappa^{*3})^{\frac{1}{1+\alpha}} \left(T^{2-\frac{2}{\kappa^*}}\right)^{-\frac{1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}\right). \\ \text{Schatten-norm: } \forall q \in [1, 2], \quad & P(\ell_{\hat{f}} - \ell_{f^*}) = O\left((R'_{max}q^*(q^* - 1))^{\frac{1}{1+\alpha}} T^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}\right). \\ \text{Graph regularizer:} \quad & P(\ell_{\hat{f}} - \ell_{f^*}) = O\left((R''_{max})^{\frac{1}{1+\alpha}} T^{\frac{-1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}\right). \end{aligned} \quad (38)$$

As mentioned earlier in Remark 7, the bounds for the class of group norm regularizers for $1 \leq q \leq 2$ is not monotonic in q ; they are minimized for $q^* = \frac{2}{3} \log T$. Therefore, we split our analysis for the group norm into two cases: first, we consider $q^* \geq \frac{2}{3} \log T$, which leads to the optimal choice $\kappa^* = q^*$, and taking the minimum of the global and local bounds gives

$$P(\ell_{\hat{f}} - \ell_{f^*}) \leq O\left(\min\left\{(R_{max}\kappa^{*3})^{\frac{1}{2}} \left(T^{2-\frac{2}{\kappa^*}}\right)^{-\frac{1}{2}} n^{-\frac{1}{2}}, (R_{max}\kappa^{*3})^{\frac{1}{1+\alpha}} \left(T^{2-\frac{2}{\kappa^*}}\right)^{-\frac{1}{1+\alpha}} n^{\frac{-\alpha}{1+\alpha}}\right\}\right).$$

It can be seen that for $\alpha \approx 1$, the minimum of the two terms are the same and local analysis has no advantages over the global one, however, for large value of α (more specially when $\alpha \rightarrow \infty$), it can be shown that local analysis improves over the global one, if T and R_{max} can grow with n such that $T/\sqrt{R_{max}} = O(\sqrt{n})$.

secondly, we assume $q^* \leq \frac{2}{3} \log T$ in which the best choice is $\kappa^* = \frac{2}{3} \log T$. Then, the excess risk bound reads as

$$P(\ell_{\hat{f}} - \ell_{f^*}) \leq O\left(\min\left\{R_{max}^{1/2}(\log T)^{3/2} T^{-1} n^{-1/2}, n^{-1}\right\}\right),$$

and the local analysis improves over the global one, when $T/\sqrt{R_{max}(\log T)^3} = O(\sqrt{n})$. A similar analysis for Schatten norm and graph regularized hypothesis spaces shows that the local analysis is beneficial over the global one whenever $T/R = O(n)$.

A close appraisal of the results in (37) and (38) points to a conservation of asymptotic rates between n and T , when all other remaining quantities are held fixed. This phenomenon is more apparent for the Schatten norm and graph-based regularization cases. It can be seen that, for both the global and local analysis results, the rates (exponents) of n and T sum up to -1 . In the local analysis case, the trade-off is determined by the value of α , which can facilitate faster n -rates and compromise with slower T -rates. A similar trade-off is witnessed in the case of group norm regularization, but this time between n and $T^{2(1-\frac{1}{\kappa^*})}$, instead of T , due to specific character of the group norm.

6.2 Comparisons to Related Works

Also, it would be interesting to compare our (global and local) results for the trace norm regularized MTL with the GRC-based excess risk bound provided in [41] wherein they apply a trace norm regularizer to capture the tasks' relatedness. It is worth mentioning that they consider a very slightly different hypothesis space for \mathbf{W} , which in our notation reads as

$$\mathcal{F}_{S_1} = \left\{ \mathbf{W} : \|\mathbf{W}\|_{S_1} \leq R'_{max} \sqrt{T} \right\}. \quad (39)$$

The intuition behind this assumption is interpreted as: assuming a common vector \mathbf{w} for all tasks, the regularizer should not be a function of number of tasks [41]. Given the task averaged covariance operator $C := 1/T \sum_{t=1}^T J_t$, the excess risk bound in [41] reads as (for the L-Lipschitz loss function ℓ , and \mathcal{F} with ranges in $[-b, b]$)

$$P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq 2LR'_{max} \left(\sqrt{\frac{\|C\|_{\infty}}{n}} + 5\sqrt{\frac{\ln(nT) + 1}{nT}} \right) + \sqrt{\frac{bLx}{nT}}.$$

One can easily verify that the trace norm is a Schatten norm with $q = 1$. Note that for any $q \geq 1$ it holds that $\mathcal{F}_{S_1} \subseteq \mathcal{F}_{S_q}$, which implies $\mathfrak{R}(\mathcal{F}_{S_1}) \leq \mathfrak{R}(\mathcal{F}_{S_q})$. This together with Theorem 17 and Theorem 16 (applied to the class of excess loss functions) can provide a GRC-based excess risk bound. Therefore, considering the trace norm hypothesis space (39) and the optimal value of $q^* = 2$, translates our global and local bounds to the following

1. GRC-based excess risk bound:

$$P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq \frac{4L}{T^{1/4}} \sqrt{\frac{R'_{max}}{nT}} \left\| (\mathbf{tr}(J_t))_{t=1}^T \right\|_1 + \sqrt{\frac{bLx}{nT}}.$$

2. LRC-based excess risk bound ($\forall \alpha > 1$):

$$P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq 256K \sqrt{\frac{\alpha + 1}{\alpha - 1}} (2dR'_{max}L^2)^{\frac{1}{1+\alpha}} B^{\frac{\alpha-1}{\alpha+1}} T^{\frac{-1}{2(1+\alpha)}} n^{\frac{-\alpha}{1+\alpha}} + \frac{(3Lb + 4BK)x}{nT}. \quad (40)$$

Now, assume that each operator J_t is of rank M and denote its maximum eigenvalue by λ_t^{max} . If $\lambda_{max} := \max_{t \in \mathbb{N}_T} \{\lambda_t^{max}\}$, then it is easy to verify that $\mathbf{tr}(J_t) \leq M\lambda_t^{max}$ and $\|C\|_{\infty} \leq \lambda_{max}$, which leads to the following GRC-based bounds

$$\text{Ours:} \quad P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq \frac{4L}{T^{1/4}} \sqrt{\frac{M\lambda_{max}R'_{max}}{n}} + \sqrt{\frac{bLx}{nT}}, \quad (41)$$

$$[41]: \quad P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq 2LR'_{max} \left(\sqrt{\frac{\lambda_{max}}{n}} + 5\sqrt{\frac{\ln(nT) + 1}{nT}} \right) + \sqrt{\frac{bLx}{nT}}. \quad (42)$$

One can observe that as $n \rightarrow \infty$, in all cases the bound approaches to zero, however our local bound in (40) at a rate of $n^{-\alpha/(1+\alpha)}$, our global bound in (41) at a slower rate of $1/\sqrt{n}$, and the one in (42) at the slowest rate of $\sqrt{\ln n/n}$.

We remark that, as $T \rightarrow \infty$, the bound in (40) and (41) vanish at a rate of $T^{-1/2(1+\alpha)}$ and $\sqrt{1/T^{1/2}}$, respectively. Also, the bound in (42) approaches to the limiting value $2LR'_{max}\sqrt{\lambda_{max}/n}$ at a faster rate of $\sqrt{\ln T/T}$, however it does not vanish even for very large value of T .

Another interesting comparison can be performed between our bounds and the one introduced in [39] for a graph regularized hypothesis spaces similar to (22). [39] provides a bound on the empirical GRC, however, similar to the proof of Corollary 12, we can easily convert it to a distribution dependent GRC bound which in our notation reads as (assuming that $\|D^{1/2}\mathbf{W}\| \leq \sqrt{T}R''_{max}$)

$$\mathfrak{R}(\mathcal{F}_G) \leq \sqrt{\frac{R''_{max}}{nT}} \left\| (D_{tt}^{-1} \mathbf{tr}(J_t))_{t=1}^T \right\|_1.$$

Now, with $\mathbf{D} = \mathbf{L} + \eta \mathbf{I}$ and the assumption of rank M for J_t s, it can be shown that

$$\begin{aligned} \left\| (\mathbf{D}_{tt}^{-1} \text{tr}(J_t))_{t=1}^T \right\|_1 &= \sum_{t=1}^T \mathbf{D}_{tt}^{-1} \text{tr}(J_t) \leq M \lambda_{\max} \left(\sum_{t=1}^T \mathbf{D}_{tt}^{-1} \right) = M \lambda_{\max} \text{tr}(\mathbf{D}^{-1}) = \\ &= M \lambda_{\max} \text{tr}(\mathbf{L} + \eta \mathbf{I})^{-1} = M \lambda_{\max} \left(\sum_{t=1}^T \frac{1}{\delta_t + \eta} + \frac{1}{\eta} \right) = M \lambda_{\max} \left(\frac{T}{\delta_{\min} + \eta} + \frac{1}{\eta} \right). \end{aligned}$$

where λ_{\max} is defined as before, and $(\delta_t)_{t=1}^T$ are the eigenvalues of Laplacian matrix \mathbf{L} with $\delta_{\min} := \min_{t \in \mathbb{N}_T} \delta_t$. Therefore, the GRC-based excess risk bounds are obtained as

$$\text{Ours:} \quad P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq \frac{2L}{\sqrt{n}} \sqrt{\frac{2M\lambda_{\max}R''_{\max}}{T^{1/2}} \left(\frac{1}{\delta_{\min}} + \frac{1}{T\eta} \right)} + \sqrt{\frac{bLx}{nT}}. \quad (43)$$

$$[39]: \quad P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq \frac{2L}{\sqrt{n}} \sqrt{M\lambda_{\max}R''_{\max} \left(\frac{1}{\delta_{\min}} + \frac{1}{T\eta} \right)} + \sqrt{\frac{bLx}{nT}}. \quad (44)$$

also, from Remark 14, the LRC-based bound is given as

$$P(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq 256K \sqrt{\frac{\alpha+1}{\alpha-1}} (dR''_{\max} L^2 \mathbf{D}_{\max}^{-1})^{\frac{1}{1+\alpha}} B^{\frac{\alpha-1}{\alpha+1}} T^{\frac{-1}{2(1+\alpha)}} n^{\frac{-\alpha}{1+\alpha}} + \frac{(3Lb + 4BK)x}{nT}. \quad (45)$$

The above results show that when $n \rightarrow \infty$, all three bounds above approach zero, albeit, the global bounds with a rate of $\sqrt{1/n}$, and the local one with a faster rate of $n^{-\alpha/\alpha+1}$, as $\alpha > 1$. Also, as $T \rightarrow \infty$, our bound in (43) vanishes at a rate of $\sqrt{1/T^{1/2}}$, while the bound in (44) approaches to the limiting value

$$\frac{2LR''_{\max}}{\sqrt{n}} \sqrt{\frac{M\lambda_{\max}}{\delta_{\min}}}$$

at a faster rate of $\sqrt{1/T}$, however the cost of learning does not vanish. At the end, our local bound in (45) approaches zero at the slowest rate of $T^{-1/2(1+\alpha)}$.

A Appendix

Proofs of the results in Sect. 2: “Talagrand-Type Inequality for Multi-Task Learning”

Our proof of Theorem 1 is based on the following Bousquet’s version of Talagrand inequality.

Theorem A.1 (Theorem 6.1 in [14]). *Let (Z, Z'_1, \dots, Z'_n) be a sequence of \mathcal{A} -measurable random variables and $(Z_k)_{k=1}^n$ be a sequence of random variables \mathcal{A}_n^k -measurable, where \mathcal{A}_n^k is a sigma field generated by $\{Z_1, \dots, Z_n\}/Z_k$. Assume that there exist $c > 0$, such that for all $k = 1, \dots, n$ the following inequalities are satisfied*

$$Z'_k \leq Z - Z_k \leq c, \quad \mathbb{E}_n^k Z'_k = 0, \quad \sum_{k=1}^n (Z - Z_k) \leq Z.$$

If we have $\sum_{k=1}^n \mathbb{E}_n^k [Z'_k]^2 \leq n\sigma^2$ and $\nu = 2c\mathbb{E}(Z) + n\sigma^2$. Then with probability at least $1 - e^{-x}$, for all $x > 0$, we have

$$Z \leq \mathbb{E}Z + \sqrt{2\nu x} + \frac{cx}{3}.$$

Proof of Theorem 1

Define the quantities

$$Z := \sup_{\mathbf{f} \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^T \frac{1}{N_t} \sum_{i=1}^{N_t} [\mathbb{E} f_t(X_t) - f_t(X_t^i)],$$

$$Z_{s,j} := \sup_{\mathbf{f} \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^T \frac{1}{N_t} \sum_{i=1}^{N_t} [\mathbb{E} f_t(X_t) - f_t(X_t^i)] - \frac{1}{TN_s} [\mathbb{E} f_s(X_s) - f_s(X_s^j)].$$

Also, let $\hat{\mathbf{f}} := (\hat{f}_1, \dots, \hat{f}_T)$ be such that $Z = \frac{1}{T} \sum_{t=1}^T \frac{1}{N_t} \sum_{i=1}^{N_t} [\mathbb{E} \hat{f}_t(X_t) - \hat{f}_t(X_t^i)]$ and similarly $\hat{\mathbf{f}}^{s,j} = (\hat{f}_1^{s,j}, \dots, \hat{f}_T^{s,j})$ be the function achieving the supremum in the definition of $Z_{s,j}$. Based on the definition of supremum, taking $n := \min_{t \in \mathbb{N}_T} N_t$, one can easily verify that

$$\frac{1}{N_s T} [\mathbb{E} \hat{f}_s^{s,j}(X_s) - \hat{f}_s^{s,j}(X_s^j)] \leq Z - Z_{s,j} \leq \frac{1}{N_s T} [\mathbb{E} \hat{f}_s(X_s) - \hat{f}_s(X_s^j)] \leq \frac{2b}{nT}.$$

Let $Z'_{s,j} := \frac{1}{N_s T} [\mathbb{E} \hat{f}_s^{s,j}(X_s) - \hat{f}_s^{s,j}(X_s^j)]$ and $c := \frac{2b}{nT}$. One can verify that Z is sub-additive, that is $\sum_{s=1}^T \sum_{j=1}^{N_s} (Z - Z_{s,j}) \leq Z$, and $\mathbb{E} Z'_{s,j} = 0$. Thus, all the conditions of Theorem A.1 are satisfied. Also, we have

$$\mathbb{E} [Z'_{s,j}]^2 = \frac{1}{N_s^2 T^2} \mathbb{E} [\mathbb{E} \hat{f}_s^{s,j}(X_s) - \hat{f}_s^{s,j}(X_s^j)]^2 \leq \frac{1}{N_s^2 T^2} \mathbb{E} [\hat{f}_s^{s,j}(X_s^j)]^2. \quad (\text{A.1})$$

Therefore,

$$\begin{aligned} \sum_{s=1}^T \sum_{j=1}^{N_s} \mathbb{E} [Z'_{s,j}]^2 &\leq \sum_{s=1}^T \sum_{j=1}^{N_s} \frac{1}{N_s^2 T^2} \mathbb{E} [\hat{f}_s^{s,j}(X_s^j)]^2 \\ &\leq \sum_{s=1}^T \sum_{j=1}^{N_s} \frac{1}{N_s^2 T^2} \sup_{\mathbf{f} \in \mathcal{F}} \mathbb{E} [f_s(X_s^j)]^2 \\ &= \sum_{s=1}^T \sum_{j=1}^{N_s} \frac{1}{N_s^2 T^2} \sup_{\mathbf{f} \in \mathcal{F}} \mathbb{E} [f_s(X_s^1)]^2 \\ &\leq \frac{r}{nT}, \end{aligned}$$

where in the equality above, we used the fact that for fixed s , X_s^j s are identically distributed. Therefore, applying Theorem A.1 for Z implies the following with probability at least $1 - e^{-x}$

$$Z \leq \mathbb{E}Z + \sqrt{2x \left(\frac{r}{nT} + 2c\mathbb{E}Z \right)} + \frac{cx}{3}.$$

Substituting $c = \frac{2b}{nT}$, and using the simple inequalities $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ and $2\sqrt{uv} \leq u+v$ for any $u, v \geq 0$, we get

$$Z \leq 2\mathbb{E}Z + \sqrt{\frac{2xr}{nT}} + \frac{8bx}{3nT}. \quad (\text{A.2})$$

The first term in the right-hand side of the above inequality, $\mathbb{E}Z$, can also be upper-bounded using the same approach as in Theorem 16 in [38]. Let X' be an *i.i.d.* copy of the X^{nT} -valued random variable X . Then,

$$\begin{aligned} \mathbb{E}Z &= \mathbb{E}_X \left[\sup_{\mathbf{f} \in \mathcal{F}} \frac{1}{T} \mathbb{E}_{X'} \left[\sum_{t=1}^T \frac{1}{N_t} \sum_{i=1}^{N_t} (f_t(X_t'^i) - f_t(X_t^i)) \right] \right] \\ &\leq \mathbb{E}_{XX'} \left[\sup_{\mathbf{f} \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^T \frac{1}{N_t} \sum_{i=1}^{N_t} (f_t(X_t'^i) - f_t(X_t^i)) \right] \\ &= \mathbb{E}_{XX'} \left[\sup_{\mathbf{f} \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^T \frac{1}{N_t} \sum_{i=1}^{N_t} \sigma_t^i (f_t(X_t'^i) - f_t(X_t^i)) \right], \end{aligned}$$

since for any i and t , the random variable $f_t(X_t'^i) - f_t(X_t^i)$ has a symmetric distribution w.r.t. 0, therefore

$$\mathbb{E}Z \leq \mathbb{E}_X \mathbb{E}_\sigma \left[\sup_{\mathbf{f} \in \mathcal{F}} \frac{2}{T} \sum_{t=1}^T \frac{1}{N_t} \sum_{i=1}^{N_t} \sigma_t^i f_t(X_t^i) \right] = 2\mathfrak{R}(\mathcal{F}),$$

where $\mathfrak{R}(\mathcal{F})$ is the Rademacher complexity of function class \mathcal{F} . Therefore, upper-bounding (A.2), one can obtain with probability at least $1 - e^{-x}$,

$$Z \leq 4\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{2xr}{nT}} + \frac{8bx}{3nT},$$

which completes the proof.

Proofs of the results in Sect. 3: “Excess MTL Risk Bounds based on Local Rademacher Complexities”

The following theorem is the core of the proof of Theorem 2 in Sect. 3.

Theorem A.2 (Distribution-dependent bound for MTL). *Let $\mathcal{F} = \{\mathbf{f} := (f_1, \dots, f_T)\}$ be a class of vector-valued functions satisfying $\sup_{t,x} |f_t(x)| \leq b$. Let $X := (X_t^i, Y_t^i)_{(t,i)=(1,1)}^{(T,n)}$ be a vector of nT independent random variables where $(X_t^1, Y_t^1), \dots, (X_t^n, Y_t^n), \forall t$ are identically distributed. Assume that there exist a constant B and a function $T : \mathcal{F} \mapsto \mathbb{R}^+$ such that for every $\mathbf{f} \in \mathcal{F}$, it holds that $P\mathbf{f}^2 \leq V(\mathbf{f}) \leq BP\mathbf{f}$. Let ψ be a sub-root function with the fixed point r^* . If ψ satisfies, for any $r \geq r^*$,*

$$B\mathfrak{R}(\mathcal{F}, r) \leq \psi(r),$$

where $\mathfrak{R}(\mathcal{F}, r)$ is the LRC of the function class \mathcal{F} defined as

$$\mathfrak{R}(\mathcal{F}, r) := \mathbb{E} \left[\sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ V(\mathbf{f}) \leq r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \right]. \quad (\text{A.3})$$

Then,

1. For any function class \mathcal{F} , $K > 1$ and $x > 0$, with probability at least $1 - e^{-x}$,

$$P\mathbf{f} \leq \frac{K}{K-1}P_n\mathbf{f} + \frac{500K}{B}r^* + \frac{(6b+10BK)x}{nT}. \quad (\text{A.4})$$

2. For any convex function class \mathcal{F} , $K > 1$ and $x > 0$, with probability at least $1 - e^{-x}$,

$$P\mathbf{f} \leq \frac{K}{K-1}P_n\mathbf{f} + \frac{32K}{B}r^* + \frac{(3b+4BK)x}{nT}. \quad (\text{A.5})$$

Proof. Define the rescaled version of \mathcal{F} (its restriction to variance radius r) as

$$\mathcal{F}_r := \left\{ \mathbf{f}' = (f'_1, \dots, f'_T), f'_t := \frac{r}{\max(r, V(\mathbf{f}))} f_t, \forall t, f_t \in \mathcal{F} \right\}. \quad (\text{A.6})$$

First, we will show that every $\mathbf{f}' \in \mathcal{F}_r$ satisfies $P\mathbf{f}'^2 \leq r$. Indeed, if we consider $\mathbf{f} \in \mathcal{F}$ such that $V(\mathbf{f}) \leq r$, then by the definition of \mathcal{F}_r , $f'_t = f_t$, hence $P\mathbf{f}'^2 = P\mathbf{f}^2 \leq V(\mathbf{f}) \leq r$. Otherwise, if $V(\mathbf{f}) \geq r$, then $f'_t = rf_t/V(\mathbf{f})$. Thus we have

$$P\mathbf{f}'^2 = \frac{1}{T} \sum_{t=1}^T P f_t'^2 = \frac{r^2}{(V(\mathbf{f}))^2} \left(\frac{1}{T} \sum_{t=1}^T P f_t^2 \right) = \frac{r^2}{(V(\mathbf{f}))^2} P\mathbf{f}^2 \leq \frac{r^2}{(V(\mathbf{f}))^2} V(\mathbf{f}) \leq r.$$

Therefore, we can conclude that for any $\mathbf{f}' \in \mathcal{F}_r$, it holds $P\mathbf{f}'^2 \leq r$. Also, as the functions in \mathcal{F} has ranges in $[-b, b]$ and $0 \leq r/\max(r, V(\mathbf{f})) \leq 1$, it can be seen that any $\mathbf{f}' \in \mathcal{F}_r$ satisfies $|\mathbf{f}'| \leq 2b$, and $\|\mathbf{f}' - P\mathbf{f}'\|_\infty \leq 2b$, consequently. Applying Theorem 1 on function class \mathcal{F}_r , for all $x > 0$, with probability greater than $1 - e^{-x}$, gives

$$\sup_{\mathbf{f}' \in \mathcal{F}_r} [P\mathbf{f}' - P_n\mathbf{f}'] \leq 4\mathfrak{R}(\mathcal{F}_r) + \sqrt{\frac{2xr}{nT}} + \frac{8bx}{3nT}. \quad (\text{A.7})$$

Now for the proof of the first part, let $\mathcal{F}(u, v) := \{\mathbf{f} : u \leq V(\mathbf{f}) \leq v\}$, and define for class \mathcal{F}_r ,

$$\mathfrak{R}_n\mathbf{f} := \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i), \quad \mathfrak{R}_n(\mathcal{F}) := \sup_{\mathbf{f} \in \mathcal{F}} \mathfrak{R}_n\mathbf{f}. \quad (\text{A.8})$$

Clearly, the Rademacher complexity of \mathcal{F}_r is $\mathbb{E}\mathfrak{R}_n(\mathcal{F}_r)$. Also, it can be shown for any sets \mathcal{A} and \mathcal{B}

$$\mathbb{E} \left[\sup_{\mathbf{f}' \in \mathcal{A} \cup \mathcal{B}} \mathfrak{R}_n\mathbf{f}' \right] \leq \mathbb{E} \left[\sup_{\mathbf{f}' \in \mathcal{A}} \mathfrak{R}_n\mathbf{f}' \right] + \mathbb{E} \left[\sup_{\mathbf{f}' \in \mathcal{B}} \mathfrak{R}_n\mathbf{f}' \right]. \quad (\text{A.9})$$

Note that for every $\mathbf{f} \in \mathcal{F}$ it holds that $V(\mathbf{f}) \leq BP\mathbf{f} \leq Bb$. Also, let $\lambda > 1$ and define k to be the smallest integer such that $r\lambda^{k+1} \geq Bb$. Thus by (A.9),

$$\begin{aligned} \mathfrak{R}(\mathcal{F}_r) &= \mathbb{E} \left[\sup_{\mathbf{f}' \in \mathcal{F}_r} \mathfrak{R}_n\mathbf{f}' \right] = \mathbb{E} \left[\sup_{\mathbf{f} \in \mathcal{F}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \frac{r}{\max(r, V(\mathbf{f}))} \sigma_t^i f_t(X_t^i) \right] \\ &\leq \mathbb{E} \left[\sup_{\mathbf{f} \in \mathcal{F}(0, r)} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \right] + \mathbb{E} \left[\sup_{\mathbf{f} \in \mathcal{F}(r, bB)} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \frac{r}{V(\mathbf{f})} \sigma_t^i f_t(X_t^i) \right] \\ &\leq \mathfrak{R}(\mathcal{F}, r) + \sum_{j=0}^k \lambda^{-j} \mathbb{E} \left[\sup_{\mathbf{f} \in \mathcal{F}(r\lambda^j, r\lambda^{j+1})} \mathfrak{R}_n\mathbf{f} \right], \end{aligned} \quad (\text{A.10})$$

where in the last inequality, we applied (A.9) for the union of intervals $\mathcal{A}_j := (r\lambda^j, r\lambda^{j+1})$. Therefore, it can be seen that $\psi(r) \geq B\mathfrak{R}(\mathcal{F}, r)$ implies

$$\begin{aligned}\mathfrak{R}(\mathcal{F}_r) &\leq \mathfrak{R}(\mathcal{F}, r) + \sum_{j=0}^k \lambda^{-j} \mathbb{E} \left[\sup_{\mathbf{f} \in \mathcal{F}(r\lambda^j, r\lambda^{j+1})} \mathfrak{R}_n \mathbf{f} \right] \\ &\leq \frac{\psi(r)}{B} + \frac{1}{B} \sum_{j=0}^k \lambda^{-j} \psi(r\lambda^{j+1}).\end{aligned}\tag{A.11}$$

Now since $\psi(r)$ is a sub-root function, by the assumption it follows that for any $\lambda \geq 1$, $\psi(\lambda r) \leq \sqrt{\lambda} \psi(r)$, hence

$$\mathfrak{R}(\mathcal{F}_r) \leq \frac{\psi(r)}{B} \left(1 + \sqrt{\lambda} \sum_{j=0}^k \lambda^{-j/2} \right).$$

By the optimal $\lambda = 4$, the right-hand side can be upper-bounded by $5\psi(r)/B$. Finally, for $r \geq r^*$, it holds $\psi(r) \leq \sqrt{r/r^*} \psi(r^*) = \sqrt{rr^*}$ and thus

$$\mathfrak{R}(\mathcal{F}_r) \leq \frac{5}{B} \sqrt{rr^*}.\tag{A.12}$$

Combining (A.7) and (A.12), for any $r \geq r^*$ and $x > 0$, with probability at least $1 - e^{-x}$, we have:

$$\sup_{\mathbf{f}' \in \mathcal{F}_r} P\mathbf{f}' - P_n\mathbf{f}' \leq \frac{20}{B} \sqrt{rr^*} + \sqrt{\frac{2xr}{nT}} + \frac{8bx}{3nT}.\tag{A.13}$$

Finally, in the following we need to convert the upper bound for functions in the weighted class \mathcal{F}_r into a bound for functions in the initial class \mathcal{F} . Denote the left hand side of the above inequality by V_r^+ . We will show that if $V_r^+ \leq \frac{r}{BK}$, then

$$P\mathbf{f} \leq \frac{K}{K-1} P_n\mathbf{f} + \frac{r}{BK}.\tag{A.14}$$

First, note that for any $\mathbf{f}' \in \mathcal{F}_r$, it holds that $P\mathbf{f}' \leq P_n\mathbf{f}' + V_r^+$. We also showed earlier, if $V(\mathbf{f}) \leq r$ then $\mathbf{f}' = \mathbf{f}$, hence

$$P\mathbf{f} \leq P_n\mathbf{f} + \frac{r}{BK}.\tag{A.15}$$

Otherwise, if $V(\mathbf{f}) \geq r$, then $\mathbf{f}' = r\mathbf{f}/V(\mathbf{f})$. Thus,

$$\frac{r}{V(\mathbf{f})} P\mathbf{f} \leq \frac{r}{V(\mathbf{f})} P_n\mathbf{f} + V_r^+ \leq \frac{r}{V(\mathbf{f})} P_n\mathbf{f} + \frac{r}{BK},\tag{A.16}$$

which coupled with $V(\mathbf{f}) \leq BP\mathbf{f}$, implies that

$$P\mathbf{f} \leq P_n\mathbf{f} + \frac{P\mathbf{f}}{K}.\tag{A.17}$$

Combining (A.15) and (A.17) implies that if $\sup_{\mathbf{f}' \in \mathcal{F}_r} P\mathbf{f}' - P_n\mathbf{f}' \leq \frac{r}{BK}$, then (A.14) holds. Setting $A = (20\sqrt{r^*}/B + \sqrt{2x/nT})$ and $C = 8bx/3nT$, the upper bound (A.13) can be written as $A\sqrt{r} + C$. Now, we want to choose $r_0 \geq r^*$ such that the upper bound of (A.13) becomes of a form r_0/BK . We achieve this by considering the largest solution of $A\sqrt{r_0} + C = r_0/BK$ which satisfies $r_0 \leq (BK)^2 A^2 + 2BKC$. Therefore, for every $\mathbf{f} \in \mathcal{F}$ we have

$$\begin{aligned}P\mathbf{f} &\leq \frac{K}{K-1} P_n\mathbf{f} + \frac{r_0}{BK} \\ &\leq \frac{K}{K-1} P_n\mathbf{f} + BKA^2 + 2C \\ &\leq \frac{K}{K-1} P_n\mathbf{f} + BK \left(\frac{400}{B^2} r^* + \frac{40}{B} \sqrt{\frac{2xr^*}{nT}} + \frac{2x}{nT} \right) + \frac{16bx}{3nT}.\end{aligned}\tag{A.18}$$

Using the fact that for any $u, v \geq 0$ and $\alpha > 0$ it holds $2\sqrt{uv} \leq \alpha u + v/\alpha$, we have $\sqrt{2xr^*/nT} \leq Bx/(5nT) + 5r^*/(2B)$, we can complete the proof.

Also, the proof of the second part follows from the fact that $\mathcal{F}_r \subseteq \{\mathbf{f} \in \text{star}(\mathcal{F}, 0) : V(\mathbf{f}) \leq r\}$. From the other side, any convex function class \mathcal{F} is star-shaped around any of its points. Therefore, $\mathfrak{R}(\mathcal{F}_r)$ in (A.7) can be bounded as $\mathfrak{R}(\mathcal{F}_r) \leq \mathcal{R}(\mathcal{F}, r) \leq \psi(r)/B$. Then, following the same argument to convert the bound for the weighted class \mathcal{F}_r into a bound for the functions in \mathcal{F} completes the proof. \square

The following lemma which is a consequence of Corollary 2.2 from [7] is essential component of the proof in Theorem 4.

Lemma 1. Assume \mathcal{F} is a class of vector-valued functions that map \mathcal{X} into $[-b, b]$ with $b > 0$. For every $x > 0$, if r satisfies

$$r \geq 16L^2b\mathbb{E}_{\sigma, X} \left\{ \sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ L^2P(\mathbf{f} - \mathbf{f}^*)^2 \leq r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \right\} + \frac{8L^2b^2x}{nT},$$

then, with probability at least $1 - e^{-x}$,

$$\left\{ \mathbf{f} \in \mathcal{F} : L^2P(\mathbf{f} - \mathbf{f}^*)^2 \leq r \right\} \subset \left\{ \mathbf{f} \in \mathcal{F} : L^2P_n(\mathbf{f} - \mathbf{f}^*)^2 \leq 2r \right\}.$$

Proof. First, define

$$\mathcal{F}_r^* := \left\{ \mathbf{f}' = (f'_1, \dots, f'_T) : \forall t, f'_t = (f_t - f_t^*)^2, f_t \in \mathcal{F}, L^2P(\mathbf{f} - \mathbf{f}^*)^2 \leq r \right\}.$$

Note that for all $t \in \mathbb{N}_T$, $(f_t - f_t^*)^2 \in [0, b^2]$. Also, for any function in \mathcal{F}_r^* , it holds that

$$P\mathbf{f}'^2 = \frac{1}{T} \sum_{t=1}^T P f_t'^2 = \frac{1}{T} \sum_{t=1}^T P (f_t - f_t^*)^4 \leq \frac{b^2}{T} \sum_{t=1}^T P (f_t - f_t^*)^2 = b^2 P(\mathbf{f} - \mathbf{f}^*)^2 \leq \frac{b^2 r}{L^2}.$$

Therefore, by Theorem 1, with probability at least $1 - e^{-x}$, every $\mathbf{f}' \in \mathcal{F}_r^*$ satisfies

$$P_n \mathbf{f}' \leq P \mathbf{f}' + 4\mathfrak{R}(\mathcal{F}_r^*) + \sqrt{\frac{2b^2xr}{nTL^2}} + \frac{8b^2x}{3nT}, \quad (\text{A.19})$$

where

$$\begin{aligned} \mathfrak{R}(\mathcal{F}_r^*) &= \mathbb{E}_{\sigma, X} \left\{ \sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ L^2P(\mathbf{f} - \mathbf{f}^*)^2 \leq r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i (f_t(X_t^i) - f_t^*(X_t^i))^2 \right\} \\ &\leq 2b\mathbb{E}_{\sigma, X} \left\{ \sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ L^2P(\mathbf{f} - \mathbf{f}^*)^2 \leq r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \right\}. \end{aligned} \quad (\text{A.20})$$

The last inequality follows from the facts that $g(x) = x^2$ is $2b$ -Lipschitz on $[-b, b]$ and \mathbf{f} is fixed. This together with (A.19), gives

$$\begin{aligned} P_n \mathbf{f}' &\leq P \mathbf{f}' + 8b\mathbb{E}_{\sigma, X} \left\{ \sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ L^2P(\mathbf{f} - \mathbf{f}^*)^2 \leq r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \right\} + \sqrt{\frac{2b^2xr}{nTL^2}} + \frac{8b^2x}{3nT} \\ &\leq \frac{r}{L^2} + 8b\mathbb{E}_{\sigma, X} \left\{ \sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ L^2P(\mathbf{f} - \mathbf{f}^*)^2 \leq r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \right\} + \frac{r}{2L^2} + \frac{11b^2x}{3nT}. \end{aligned} \quad (\text{A.21})$$

Note that multiplying both side by L^2 completes the proof. \square

proof of Theorem 4

Define the function $\psi(r)$ as

$$\psi(r) = \frac{c_1}{2} \mathbb{E} \left[\sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ L^2 P(\mathbf{f} - \mathbf{f}^*)^2 \leq r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \right] + \frac{(c_2 - c_1)x}{nT}. \quad (\text{A.22})$$

Since \mathcal{F} is convex, it is star-shaped around any of its points, thus using Lemma 3.4 in [7] it can be shown that $\psi(r)$ defined in (A.22) is a sub-root function. with the help of Corollary 3 and Assumptions 1, we have with probability at least $1 - e^{-x}$

$$L^2 P(\hat{\mathbf{f}} - \mathbf{f}^*)^2 \leq BP(\ell_{\hat{\mathbf{f}}} - \ell_{\mathbf{f}^*}) \leq 32Kr + \frac{(3Lb + 4BK)Bx}{nT}. \quad (\text{A.23})$$

Denote the right hand side of the above inequality by s . Since $s \geq r \geq r^*$, then by the property of sub-root functions it holds that $s \geq \psi(s)$, and thus

$$s \geq 16L^2 b \mathbb{E} \left[\sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ L^2 P(\mathbf{f} - \mathbf{f}^*)^2 \leq r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \right] + \frac{8L^2 b^2 x}{nT}.$$

Applying Lemma 1, we have with probability at least $1 - e^{-x}$,

$$\left\{ \mathbf{f} \in \mathcal{F}, L^2 P(\mathbf{f} - \mathbf{f}^*)^2 \leq s \right\} \subset \left\{ \mathbf{f} \in \mathcal{F}, L^2 P_n(\mathbf{f} - \mathbf{f}^*)^2 \leq 2s \right\}.$$

Combining this with (A.23), gives with probability at least $1 - 2e^{-x}$,

$$\begin{aligned} L^2 P_n(\hat{\mathbf{f}} - \mathbf{f}^*)^2 &\leq 2 \left(32Kr + \frac{(3Lb + 4BK)Bx}{nT} \right) \\ &\leq 2 \left(32K + \frac{(3Lb + 4BK)B}{c_2} \right) r. \end{aligned} \quad (\text{A.24})$$

where in the last inequality we used the fact that $r \geq \psi(r) \geq c_2 x / nT$. Taking $c = 2(32K + (3Lb + 4BK)B/c_2)$, and applying triangle inequality, if (A.24) holds, then for any $\mathbf{f} \in \mathcal{F}$, we have

$$\begin{aligned} L^2 P_n(\mathbf{f} - \hat{\mathbf{f}})^2 &\leq \left(\sqrt{L^2 P_n(\mathbf{f} - \mathbf{f}^*)^2} + \sqrt{L^2 P_n(\mathbf{f}^* - \hat{\mathbf{f}})^2} \right)^2 \\ &\quad \left(\sqrt{L^2 P_n(\mathbf{f} - \mathbf{f}^*)^2} + \sqrt{cr} \right)^2. \end{aligned} \quad (\text{A.25})$$

Now, applying Lemma 1 for $r \geq \psi(r)$, implies that with probability at least $1 - 3e^{-x}$,

$$\left\{ \mathbf{f} \in \mathcal{F}, L^2 P(\mathbf{f} - \mathbf{f}^*)^2 \leq r \right\} \subset \left\{ \mathbf{f} \in \mathcal{F}, L^2 P_n(\mathbf{f} - \mathbf{f}^*)^2 \leq 2r \right\},$$

which coupled with (A.25), implies that with probability at least $1 - 3e^{-x}$,

$$\left\{ \mathbf{f} \in \mathcal{F}, L^2 P(\mathbf{f} - \mathbf{f}^*)^2 \leq r \right\} \subset \left\{ \mathbf{f} \in \mathcal{F}, L^2 P_n(\mathbf{f} - \hat{\mathbf{f}})^2 \leq (\sqrt{2} + \sqrt{c})^2 r \right\}.$$

Thus, with the help of Lemma A.4 in [7], it can be shown that with probability at least $1 - 4e^{-x}$,

$$\begin{aligned}
\psi(r) &\leq c_1 \mathbb{E}_\sigma \left[\sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ L^2 P(\mathbf{f} - \mathbf{f}^*)^2 \leq r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \middle| \{x_t^i\}_{t \in \mathbb{N}_T, i \in \mathbb{N}_n} \right] + \frac{c_2 x}{nT} \\
&\leq c_1 \mathbb{E}_\sigma \left[\sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ L^2 P_n(\mathbf{f} - \hat{\mathbf{f}})^2 \leq (\sqrt{2} + \sqrt{c})^2 r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \middle| \{x_t^i\}_{t \in \mathbb{N}_T, i \in \mathbb{N}_n} \right] + \frac{c_2 x}{nT} \\
&\leq c_1 \mathbb{E}_\sigma \left[\sup_{\substack{\mathbf{f} \in \mathcal{F}, \\ \forall t, L^2 P_n(\mathbf{f} - \hat{\mathbf{f}})^2 \leq (4+2c)r}} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \middle| \{x_t^i\}_{t \in \mathbb{N}_T, i \in \mathbb{N}_n} \right] + \frac{c_2 x}{nT} \\
&\leq \hat{\psi}(r).
\end{aligned} \tag{A.26}$$

Setting $r = r^*$ and applying Lemma 4.3 of [7], gives $r^* \leq \hat{r}^*$ which together with (A.23) yields the result.

Proofs of the results in Sect. 4: “Local Rademacher Complexity Bounds for MTL models with Strongly Convex Regularizers”

In the following, we would like to provide some basic notions of convex analysis which are helpful in understanding the results of Sect. 4.

Definition 1 (Strong Convexity). *A function $R : \mathcal{X} \mapsto \mathbb{R}$ is μ -strong convex w.r.t. a norm $\|\cdot\|$ if and only if $\forall x, y \in \mathcal{X}$ and $\forall \alpha \in (0, 1)$, we have*

$$R(\alpha x + (1 - \alpha)y) \leq \alpha R(x) + (1 - \alpha)R(y) - \frac{\mu}{2} \alpha(1 - \alpha) \|x - y\|^2.$$

Definition 2 (Strong Smoothness). *A function $R^* : \mathcal{X} \mapsto \mathbb{R}$ is $\frac{1}{\mu}$ -strong smooth w.r.t. a norm $\|\cdot\|_*$ if and only if R^* is everywhere differentiable and $\forall x, y \in \mathcal{X}$, we have*

$$R^*(x + y) \leq R^*(x) + \langle \nabla R^*(x), y \rangle + \frac{1}{2\mu} \|y\|_*^2.$$

Property 1 (Theorem 3 in [21]: Strong convexity/strong smoothness duality). *A function R is μ -strongly convex w.r.t. the norm $\|\cdot\|$ if and only if its Fenchel conjugate R^* is $\frac{1}{\mu}$ -strongly smooth w.r.t. the dual norm $\|\cdot\|_*$. The Fenchel conjugate R^* is defined as*

$$R^*(\mathbf{w}) := \sup_{\mathbf{v}} \{ \langle \mathbf{w}, \mathbf{v} \rangle - R(\mathbf{v}) \}.$$

Property 2 (Fenchel-Young inequality). *The definition of Fenchel dual implies that for any strong convex function R ,*

$$\forall \mathbf{w}, \mathbf{v} \in S, \langle \mathbf{w}, \mathbf{v} \rangle \leq R(\mathbf{w}) + R^*(\mathbf{v}).$$

Combining this with the strong duality property of R^* gives the following

$$\langle \mathbf{w}, \mathbf{v} \rangle - R(\mathbf{w}) \leq R^*(\mathbf{v}) \leq R^*(\mathbf{0}) + \langle \nabla R^*(\mathbf{0}), \mathbf{v} \rangle + \frac{1}{2\mu} \|\mathbf{v}\|_*^2. \tag{A.27}$$

Lemma 2. *Assume that the conditions of Theorem 5 hold. Then, for ever $\mathbf{f} \in \mathcal{F}_q$,*

$$\begin{aligned}
(a) \quad & P \mathbf{f}^2 \leq r \text{ implies } 1/T \sum_{t=1}^T \sum_{j=1}^\infty \lambda_t^j \left\langle \mathbf{w}_t, \mathbf{u}_t^j \right\rangle^2 \leq r. \\
(b) \quad & \mathbb{E}_{X, \sigma} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \mathbf{u}_t^j \right\rangle^2 = \frac{\lambda_t^j}{n}.
\end{aligned}$$

Proof.

Part (a)

$$\begin{aligned}
P\mathbf{f}^2 &= \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left(\langle \mathbf{w}_t, \phi(X_t^i) \rangle \right)^2 \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left(\langle \mathbf{w}_t \otimes \mathbf{w}_t, \phi(X_t^i) \otimes \phi(X_t^i) \rangle \right) \\
&= \frac{1}{T} \sum_{t=1}^T \langle \mathbf{w}_t \otimes \mathbf{w}_t, \mathbb{E}_X (\phi(X_t^i) \otimes \phi(X_t^i)) \rangle = \frac{1}{T} \sum_{t=1}^T \sum_{j=1}^{\infty} \lambda_t^j \langle \mathbf{w}_t \otimes \mathbf{w}_t, \mathbf{u}_t^j \otimes \mathbf{u}_t^j \rangle \\
&= \frac{1}{T} \sum_{t=1}^T \sum_{j=1}^{\infty} \lambda_t^j \langle \mathbf{w}_t, \mathbf{u}_t^j \rangle \langle \mathbf{w}_t, \mathbf{u}_t^j \rangle = \frac{1}{T} \sum_{t=1}^T \sum_{j=1}^{\infty} \lambda_t^j \langle \mathbf{w}_t, \mathbf{u}_t^j \rangle^2 \leq r.
\end{aligned}$$

Part (b)

$$\begin{aligned}
\mathbb{E}_{X,\sigma} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \mathbf{u}_t^j \right\rangle^2 &= \frac{1}{n^2} \mathbb{E}_{X,\sigma} \sum_{i,k=1}^n \sigma_t^i \sigma_t^k \langle \phi(X_t^i), \mathbf{u}_t^j \rangle \langle \phi(X_t^k), \mathbf{u}_t^j \rangle \\
&\stackrel{\sigma_t \text{ i.i.d.}}{=} \frac{1}{n^2} \mathbb{E}_X \left(\sum_{i=1}^n \langle \phi(X_t^i), \mathbf{u}_t^j \rangle^2 \right) = \frac{1}{n} \left\langle \frac{1}{n} \sum_{i=1}^n \mathbb{E}_X (\phi(X_t^i) \otimes \phi(X_t^i)), \mathbf{u}_t^j \otimes \mathbf{u}_t^j \right\rangle \\
&= \frac{1}{n} \sum_{l=1}^{\infty} \lambda_t^l \langle \mathbf{u}_t^l \otimes \mathbf{u}_t^l, \mathbf{u}_t^j \otimes \mathbf{u}_t^j \rangle = \frac{\lambda_t^j}{n}.
\end{aligned}$$

□

We will use following lemmas in the proof of the LRC bound for the $L_{2,q}$ -group norm regularized MTL in Corollary 6.

Lemma 3 (Khintchine-Kahane Inequality [47]). *Let \mathcal{H} be an inner-product space with induced norm $\|\cdot\|_{\mathcal{H}}$, $v_1, \dots, v_M \in \mathcal{H}$ and $\sigma_1, \dots, \sigma_n$ i.i.d. Rademacher random variables. Then, for any $p \geq 1$, we have that*

$$\mathbb{E}_{\sigma} \left\| \sum_{i=1}^n \sigma_i v_i \right\|_{\mathcal{H}}^p \leq \left(c \sum_{i=1}^n \|v_i\|_{\mathcal{H}}^2 \right)^{\frac{p}{2}}. \quad (\text{A.28})$$

where $c := \max\{1, p-1\}$. The inequality also holds for p in place of c .

Lemma 4 (Rosenthal-Young Inequality; Lemma 3 of [24]). *Let X_1, \dots, X_n be independent, non-negative random variables satisfying $X_i \leq B < +\infty$ almost surely for all $i = 1, \dots, n$. If $q \geq \frac{1}{2}$, $c_q := (2qe)^q$, then it holds*

$$\mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^q \leq C_q \left[\left(\frac{B}{n} \right)^q + \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E} X_i \right)^q \right]. \quad (\text{A.29})$$

Proof of Corollary 6

For the group norm regularizer $\|\mathbf{W}\|_{2,q}$, we can further bound the expectation term in (12) for $\mathbf{D} = \mathbf{I}$ as following

$$\begin{aligned}
\mathbb{E} &:= \mathbb{E}_{X,\sigma} \left\| \left(\sum_{j>h_t} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right)_{t=1}^T \right\|_{2,q^*}^2 \\
&= \mathbb{E}_{X,\sigma} \left(\sum_{t=1}^T \left\| \sum_{j>h_t} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right\|^{q^*} \right)^{\frac{2}{q^*}} \\
&\stackrel{\text{Jensen}}{\leq} \mathbb{E}_X \left(\sum_{t=1}^T \mathbb{E}_\sigma \left\| \sum_{j>h_t} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right\|^{q^*} \right)^{\frac{2}{q^*}} \\
&\stackrel{(\text{A.28})}{\leq} \mathbb{E}_X \left(\sum_{t=1}^T \left(q^* \sum_{i=1}^n \left\| \sum_{j>h_t} \left\langle \frac{1}{n} \phi(X_t^i), \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right\|^2 \right)^{\frac{q^*}{2}} \right)^{\frac{2}{q^*}} \\
&= \frac{q^*}{n} \mathbb{E}_X \left(\sum_{t=1}^T \left(\sum_{j>h_t} \frac{1}{n} \sum_{i=1}^n \left\langle \phi(X_t^i), \mathbf{u}_t^j \right\rangle^2 \right)^{\frac{q^*}{2}} \right)^{\frac{2}{q^*}} \\
&\stackrel{\text{Jensen}}{\leq} \frac{q^*}{n} \left(\sum_{t=1}^T \mathbb{E}_X \left(\sum_{j>h_t} \frac{1}{n} \sum_{i=1}^n \left\langle \phi(X_t^i), \mathbf{u}_t^j \right\rangle^2 \right)^{\frac{q^*}{2}} \right)^{\frac{2}{q^*}} \tag{A.30}
\end{aligned}$$

Note that for $q \leq 2$, it holds that $q^*/2 \geq 1$. Therefore we cannot employ Jensen inequality to move the expectation operator inside the inner term, and we need to apply the Rosenthal-Young (R+Y) inequality (see Lemma 4 in the Appendix), which yields

$$\begin{aligned}
\mathbb{E} &\stackrel{\text{R+Y}}{\leq} \frac{q^*}{n} \left(\sum_{t=1}^T (eq^*)^{\frac{q^*}{2}} \left(\left(\frac{\beta}{n} \right)^{\frac{q^*}{2}} + \left(\sum_{j>h_t} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_X \left\langle \phi(X_t^i), \mathbf{u}_t^j \right\rangle^2 \right)^{\frac{q^*}{2}} \right) \right)^{\frac{2}{q^*}} \\
&= \frac{q^*}{n} \left(\sum_{t=1}^T (eq^*)^{\frac{q^*}{2}} \left(\left(\frac{\beta}{n} \right)^{\frac{q^*}{2}} + \left(\sum_{j>h_t} \lambda_t^j \right)^{\frac{q^*}{2}} \right) \right)^{\frac{2}{q^*}}. \tag{A.31}
\end{aligned}$$

This can be further bounded, using the sub-additivity of $q^*\sqrt{\cdot}$ and $\sqrt{\cdot}$ respectively in $(\dagger\dagger)$ and (\dagger) below,

$$\begin{aligned}
\mathbb{E} &\stackrel{(\dagger)}{\leq} \frac{eq^{*2}}{n} \left[\left(T \left(\frac{\beta}{n} \right)^{\frac{q^*}{2}} \right)^{\frac{2}{q^*}} + \left(\sum_{t=1}^T \left(\sum_{j>h_t} \lambda_t^j \right)^{\frac{q^*}{2}} \right)^{\frac{2}{q^*}} \right] \\
&\stackrel{(\dagger\dagger)}{\leq} \frac{eq^{*2}}{n} \left[\frac{\beta T^{\frac{2}{q^*}}}{n} + \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}} \right] \\
&= \frac{\beta eq^{*2} T^{\frac{2}{q^*}}}{n^2} + \frac{eq^{*2}}{n} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}.
\end{aligned} \tag{A.32}$$

Plugging this into (12), and using subadditivity of $\sqrt{\cdot}$ gives,

$$A_2(\mathcal{F}_q) \leq \sqrt{\frac{2eq^{*2}R_{max}}{nT^2\mu} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}^2} + \frac{\sqrt{2\beta e R_{max}} q^* T^{\frac{1}{q^*}}}{nT\sqrt{\mu}}. \tag{A.33}$$

Now, combining (11) and (A.33) provides the bound on $\mathfrak{R}(\mathcal{F}_q, r)$ as

$$\begin{aligned}
\mathfrak{R}(\mathcal{F}_q, r) &\leq \sqrt{\frac{r \sum_{t=1}^T h_t}{nT}} + \sqrt{\frac{2eq^{*2}R_{max}}{nT^2\mu} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}^2} + \frac{\sqrt{2\beta e R_{max}} q^* T^{\frac{1}{q^*}}}{nT\sqrt{\mu}} \\
&\stackrel{(\star)}{\leq} \sqrt{\frac{2}{nT} \left(r \sum_{t=1}^T h_t + \frac{2eq^{*2}R_{max}}{T\mu} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}} \right)} + \frac{\sqrt{2\beta e R_{max}} q^* T^{\frac{1}{q^*}}}{nT\sqrt{\mu}} \\
&\stackrel{(\star\star)}{\leq} \sqrt{\frac{2}{nT} \left(r T^{1-\frac{2}{q^*}} \left\| (h_t)_{t=1}^T \right\|_{\frac{q^*}{2}} + \frac{2eq^{*2}R_{max}}{T\mu} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}} \right)} + \frac{\sqrt{2\beta e R_{max}} q^* T^{\frac{1}{q^*}}}{nT\sqrt{\mu}} \\
&\stackrel{(\star\star\star)}{\leq} \sqrt{\frac{4}{nT} \left\| \left(r T^{1-\frac{2}{q^*}} h_t + \frac{2eq^{*2}R_{max}}{T\mu} \sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\beta e R_{max}} q^* T^{\frac{1}{q^*}}}{nT\sqrt{\mu}}.
\end{aligned} \tag{A.34}$$

where in (\star) , $(\star\star)$ and $(\star\star\star)$ we applied following inequalities receptively, according which for all non-negative numbers α_1 and α_2 , and non-negative vectors $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^T$ with $0 \leq q \leq p \leq \infty$ and $s \geq 1$ it holds

$$\begin{aligned}
(\star) \quad &\sqrt{\alpha_1} + \sqrt{\alpha_2} \leq \sqrt{2(\alpha_1 + \alpha_2)} \\
(\star\star) \quad &l_p - l_q - l_q : \quad \|\mathbf{a}_1\|_q = \langle \mathbf{1}, \mathbf{a}_1 \rangle^{\frac{1}{q}} \stackrel{\text{H\"older}}{\leq} \left(\|\mathbf{1}\|_{(p/q)^*} \|\mathbf{a}_1^q\|_{(p/q)} \right)^{\frac{1}{q}} = T^{\frac{1}{q} - \frac{1}{p}} \|\mathbf{a}_1\|_p \\
(\star\star\star) \quad &\|\mathbf{a}_1\|_s + \|\mathbf{a}_2\|_s \leq 2^{1-\frac{1}{s}} \|\mathbf{a}_1 + \mathbf{a}_2\|_s \leq 2 \|\mathbf{a}_1 + \mathbf{a}_2\|_s.
\end{aligned}$$

Since inequality $(\star\star\star)$ holds for all non-negative h_t , it follows

$$\begin{aligned}\mathfrak{R}(\mathcal{F}_q, r) &\leq \sqrt{\frac{4}{nT} \left\| \left(\min_{h_t \geq 0} rT^{1-\frac{2}{q^*}} h_t + \frac{2eq^{*2}R_{max}}{T\mu} \sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\beta e R_{max}} q^* T^{\frac{1}{q^*}}}{nT\sqrt{\mu}} \\ &\leq \sqrt{\frac{4}{nT} \left\| \left(\sum_{j=1}^{\infty} \min \left(rT^{1-\frac{2}{q^*}}, \frac{2eq^{*2}R_{max}}{T\mu} \lambda_t^j \right) \right)_{t=1}^T \right\|_{\frac{q^*}{2}}} + \frac{\sqrt{2\beta e R_{max}} q^* T^{\frac{1}{q^*}}}{nT\sqrt{\mu}}.\end{aligned}$$

Also, from Theorem 3 and Theorem 13 in [21], it can be shown that $R(\mathbf{W}) = \|\mathbf{W}\|_{2,q}^2$ is $\frac{1}{q^*}$ -strongly convex w.r.t. the group norm $\|\cdot\|_{2,q^*}$, we can conclude the result.

Proof of Theorem 9

$$\begin{aligned}\mathfrak{R}(\mathcal{F}_{q,R,T}, r) &= \frac{1}{T} \mathbb{E}_{X,\sigma} \left\{ \sup_{\substack{P\mathbf{f}^2 \leq r, \\ \|\mathbf{W}\|_{2,q} \leq R_{max}}} \sum_{t=1}^T \left\langle \mathbf{w}_t, \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i) \right\rangle \right\} \\ &= \frac{1}{T} \mathbb{E}_{X,\sigma} \left\{ \sup_{\substack{1/T \sum_{t=1}^T \mathbb{E} \langle \mathbf{w}_t, \phi(X_t) \rangle^2 \leq r, \\ \|\mathbf{W}\|_{2,q} \leq R_{max}}} \sum_{t=1}^T \left\langle \mathbf{w}_t, \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i) \right\rangle \right\} \\ &\geq \frac{1}{T} \mathbb{E}_{X,\sigma} \left\{ \sup_{\substack{\forall t \mathbb{E}_X \langle \mathbf{w}_t, \phi(X_t) \rangle^2 \leq r, \\ \|\mathbf{W}\|_{2,q} \leq R_{max}, \\ \|\mathbf{w}_1\|_2 = \dots = \|\mathbf{w}_t\|_2}} \sum_{t=1}^T \left\langle \mathbf{w}_t, \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i) \right\rangle \right\} \\ &= \frac{1}{T} \mathbb{E}_{X,\sigma} \left\{ \sup_{\substack{\forall t \mathbb{E}_X \langle \mathbf{w}_t, \phi(X_t) \rangle^2 \leq r, \\ \forall t \|\mathbf{w}_t\|_2 \leq R_{max} T^{-\frac{1}{q}}}} \sum_{t=1}^T \left\langle \mathbf{w}_t, \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i) \right\rangle \right\} \\ &= \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{X,\sigma} \left\{ \sup_{\substack{\forall t \mathbb{E}_X \langle \mathbf{w}_t, \phi(X_t) \rangle^2 \leq r, \\ \forall t \|\mathbf{w}_t\|_2 \leq R_{max} T^{-\frac{1}{q}}}} \left\langle \mathbf{w}_t, \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i) \right\rangle \right\} \\ &= \mathbb{E}_{X,\sigma} \left\{ \sup_{\substack{\mathbb{E}_X \langle \mathbf{w}_1, \phi(X_1) \rangle^2 \leq r, \\ \|\mathbf{w}_1\|_2 \leq R_{max} T^{-\frac{1}{q}}}} \left\langle \mathbf{w}_1, \frac{1}{n} \sum_{i=1}^n \sigma_1^i \phi(X_1^i) \right\rangle \right\} \\ &= \mathfrak{R}(\mathcal{F}_{1,RT^{-\frac{1}{q}},1}, r).\end{aligned}$$

According to [43], it can be shown that there is a constant c such that if $\lambda_t^1 \geq \frac{1}{nR_{max}^2}$, then for all $r \geq \frac{1}{n}$ it holds $\mathfrak{R}(\mathcal{F}_{1,RT^{-\frac{1}{q}},1}, r) \geq \sqrt{\frac{c}{n} \sum_{j=1}^{\infty} \min \left(r, R^2 T^{-\frac{2}{q}} \lambda_1^j \right)}$, which with some algebra manipulations gives the desired result.

The following lemma is used in the proof of the LRC bounds for the L_{S_q} -Schatten norm regularized MTL in Corollary 10.

Lemma 5 (Non-commutative Khintchine's inequality [33]). *Let $\mathbf{Q}_1, \dots, \mathbf{Q}_n$ be a set of arbitrary $m \times n$ matrices, and let $\sigma_1, \dots, \sigma_n$ be a sequence of independent Bernoulli random variables. Then for all $p \geq 2$,*

$$\left[\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i \mathbf{Q}_i \right\|_{S_p}^p \right]^{1/p} \leq p^{1/2} \max \left\{ \left\| \left(\sum_{i=1}^n \mathbf{Q}_i^T \mathbf{Q}_i \right)^{1/2} \right\|_{S_p}, \left\| \left(\sum_{i=1}^n \mathbf{Q}_i \mathbf{Q}_i^T \right)^{1/2} \right\|_{S_p} \right\}. \quad (\text{A.35})$$

Proof of Corollary 10

In order to find an LRC bound for a L_{S_q} -Schatten norm regularized hypothesis space (21), one just needs to bound the expectation term in (8). Define \mathbf{U}_t^i as a matrix with T columns where its only non-zero t^{th} column is defined as $\sum_{j>h_t} \left\langle \frac{1}{n} \phi(X_t^i), \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j$. Also, note that for the Schatten norm regularized hypothesis space (21), it holds that $\mathbf{D} = \mathbf{I}$. Therefore, applying Lemma 5 yields,

$$\begin{aligned} \mathbb{E}_{X,\sigma} \left\| \mathbf{D}^{-1/2} \mathbf{V} \right\|_*^2 &= \mathbb{E}_{X,\sigma} \left\| \left(\sum_{j>h_t} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right)_{t=1}^T \right\|_{S_{q^*}}^2 \\ &= \mathbb{E}_{X,\sigma} \left\| \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i \mathbf{U}_t^i \right\|_{S_{q^*}}^2 \stackrel{\text{Jensen}}{\leq} \mathbb{E}_X \left\{ \mathbb{E}_\sigma \left\| \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i \mathbf{U}_t^i \right\|_{S_{q^*}}^{q^*} \right\}^{\frac{2}{q^*}} \\ &\stackrel{(\text{A.35})}{\leq} \mathbb{E}_X \left\{ q^{*1/2} \max \left\{ \left\| \left(\sum_{t=1}^T \sum_{i=1}^n (\mathbf{U}_t^i)^T \mathbf{U}_t^i \right)^{1/2} \right\|_{S_{q^*}}, \left\| \left(\sum_{t=1}^T \sum_{i=1}^n \mathbf{U}_t^i (\mathbf{U}_t^i)^T \right)^{1/2} \right\|_{S_{q^*}} \right\} \right\}^2 \\ &\stackrel{\textcircled{A}}{=} q^* \mathbb{E}_X \left\| \left(\sum_{t=1}^T \sum_{i=1}^n (\mathbf{U}_t^i)^T \mathbf{U}_t^i \right)^{1/2} \right\|_{S_{q^*}}^2 = q^* \mathbb{E}_X \left(\text{tr} \left(\sum_{t=1}^T \sum_{i=1}^n (\mathbf{U}_t^i)^T \mathbf{U}_t^i \right)^{\frac{q^*}{2}} \right)^{\frac{2}{q^*}} \\ &= q^* \mathbb{E}_X \left(\left(\sum_{t=1}^T \sum_{i=1}^n \left\| \sum_{j>h_t} \left\langle \frac{1}{n} \phi(X_t^i), \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right\|^2 \right)^{\frac{q^*}{2}} \right)^{\frac{2}{q^*}} \\ &= q^* \mathbb{E}_X \left(\sum_{t=1}^T \sum_{i=1}^n \left\| \sum_{j>h_t} \left\langle \frac{1}{n} \phi(X_t^i), \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right\|^2 \right) \\ &= \frac{q^*}{n^2} \mathbb{E}_X \left(\sum_{t=1}^T \sum_{i=1}^n \sum_{j>h_t} \left\langle \phi(X_t^i), \mathbf{u}_t^j \right\rangle^2 \right) \\ &\stackrel{\text{Jensen}}{\leq} \frac{q^*}{n^2} \left(\sum_{t=1}^T \sum_{i=1}^n \sum_{j>h_t} \lambda_t^j \right) = \frac{q^*}{n} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_1. \end{aligned} \quad (\text{A.36})$$

where in \textcircled{A} , we assumed that the first term in the max argument is the largest. Note that using Theorem 11 in [21], it can be shown that the regularization function $R(\mathbf{W}) = \|\mathbf{W}\|_{S_q}^2$ with $q \in [1, 2]$ is $(q-1)$ -strongly convex w.r.t. q -Schatten norm $\|\cdot\|_{S_q}$. Plugging this into (8) completes the proof.

Proof of Corollary 12

Similar to the proof of Corollary 10, for the graph regularized hypothesis space (22), one can bound the expectation term in (8) as

$$\begin{aligned}
\mathbb{E}_{X,\sigma} \left\| \mathbf{D}^{-1/2} \mathbf{V} \right\|_*^2 &= \mathbb{E}_{X,\sigma} \left[\text{tr} \left(\mathbf{V}^T \mathbf{D}^{-1} \mathbf{V} \right) \right] \\
&\stackrel{\text{Jensen}}{\leq} \mathbb{E}_X \left(\frac{1}{n^2} \sum_{t,s=1}^{T,T} \sum_{i,l=1}^{n,n} \sum_{j>h_t} \sum_{k>h_s} \mathbf{D}_{st}^{-1} \mathbb{E}_\sigma (\sigma_t^i \sigma_s^l) \left\langle \phi(X_t^i), \mathbf{u}_t^j \right\rangle \left\langle \phi(X_s^l), \mathbf{u}_s^k \right\rangle \left\langle \mathbf{u}_t^j, \mathbf{u}_s^k \right\rangle \right) \\
&= \mathbb{E}_X \left(\frac{1}{n} \sum_{t=1}^T \mathbf{D}_{tt}^{-1} \sum_{j>h_t} \frac{1}{n} \sum_{i=1}^n \left\langle \phi(X_t^i), \mathbf{u}_t^j \right\rangle^2 \right) \\
&= \frac{1}{n} \sum_{t=1}^T \mathbf{D}_{tt}^{-1} \sum_{j>h_t} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_X \left\langle \phi(X_t^i), \mathbf{u}_t^j \right\rangle^2 \\
&= \frac{1}{n} \sum_{t=1}^T \sum_{j>h_t} \mathbf{D}_{tt}^{-1} \lambda_t^j = \frac{1}{n} \left\| \left(\mathbf{D}_{tt}^{-1} \sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_1. \tag{A.37}
\end{aligned}$$

proof of Corollary 15

First notice that $\hat{\mathfrak{R}}(\mathcal{F}_q^*, c_3 r) \leq 2\hat{\mathfrak{R}}(\mathcal{F}_q, \frac{c_3 r}{4L^2})$. Also, similar to the proof of Remark 6 it can be show that

$$\hat{\mathfrak{R}}(\mathcal{F}_q, \frac{c_3 r}{4L^2}) \leq \sqrt{\frac{c_3 r \sum_{t=1}^T h_t}{4nTL^2}} + \sqrt{\frac{2q^{*2} R_{max}}{nT^2} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}}.$$

Therefore,

$$\begin{aligned}
\hat{\psi}_n(r) &\leq 2c_1 \left(\sqrt{\frac{c_3 r \sum_{t=1}^T h_t}{4nTL^2}} + \sqrt{\frac{2q^{*2} R_{max}}{nT^2} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}} \right) + \frac{c_2 x}{nT} \\
&= \sqrt{\frac{c_1^2 c_3 r \sum_{t=1}^T h_t}{nTL^2}} + \sqrt{\frac{8c_1^2 q^{*2} R_{max}}{nT^2} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}} + \frac{c_2 x}{nT}.
\end{aligned}$$

Denote the right hand side by $\hat{\psi}_n^{ub}(r)$. Solving the fixed point equation $\hat{\psi}_n^{ub}(r) = \sqrt{\alpha r} + \gamma = r$ for

$$\alpha = \frac{c_1^2 c_3 \sum_{t=1}^T h_t}{nTL^2}, \quad \gamma = \sqrt{\frac{8c_1^2 q^{*2} R_{max}}{nT^2} \left\| \left(\sum_{j>h_t} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}} + \frac{c_2 x}{nT} \tag{A.38}$$

gives $\hat{r}^* \leq \alpha + 2\gamma$. Substituting α and γ completes the proof.

Proof of the results in Sect. 6.1: “Global vs. Local Rademacher Complexity Bounds”

Proof of Theorem 17

Note that regarding the definition of A_2 in (10), the global rademacher complexity for each case can be obtained by replacing the tail-sum $\sum_{j>h_t} \lambda_t^j$ in the bound of its corresponding $A_2(\mathcal{F})$ by $\sum_{j=1}^{\infty} \lambda_t^j = \text{tr}(J_t)$. Indeed, similar to the proof of Theorem 5, it can be shown that for the group norm with $\kappa = q \in [1, 2]$,

$$\begin{aligned} \mathfrak{R}(\mathcal{F}_q) &= \mathbb{E}_{X, \sigma} \left\{ \sup_{\mathbf{f}=(f_1, \dots, f_T) \in \mathcal{F}_q} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \sigma_t^i f_t(X_t^i) \right\} \\ &\leq \frac{1}{T} \sqrt{\frac{2R}{\mu} \mathbb{E}_{X, \sigma} \left\| \left(\frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i) \right)_{t=1}^T \right\|_{2, q^*}^2}. \end{aligned}$$

Also, one can verify the following

$$\begin{aligned} \mathbb{E}_{X, \sigma} \left\| \left(\frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i) \right)_{t=1}^T \right\|_{2, q^*}^2 &= \mathbb{E}_{X, \sigma} \left\| \left(\sum_{j=1}^{\infty} \left\langle \frac{1}{n} \sum_{i=1}^n \sigma_t^i \phi(X_t^i), \mathbf{u}_t^j \right\rangle \mathbf{u}_t^j \right)_{t=1}^T \right\|_{2, q^*}^2 \\ &\leq \frac{q^{*2} \beta e T^{\frac{2}{q^*}}}{n^2} + \frac{q^{*2} e}{n} \left\| \left(\sum_{j=1}^{\infty} \lambda_t^j \right)_{t=1}^T \right\|_{\frac{q^*}{2}}^2 \\ &= \frac{q^{*2} \beta e T^{\frac{2}{q^*}}}{n^2} + \frac{q^{*2} e}{n} \left\| (\text{tr}(J_t))_{t=1}^T \right\|_{\frac{q^*}{2}}^2. \end{aligned} \tag{A.39}$$

where the last inequality obtained in a similar way in (A.32). The GRC bounds for the other cases can be easily derived in a very similar way.

References

- [1] Qi An, Chunping Wang, Ivo Shterev, Eric Wang, Lawrence Carin, and David B Dunson. Hierarchical kernel stick-breaking process for multi-task image analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 17–24. ACM, 2008.
- [2] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6:1817–1853, 2005.
- [3] Andreas Argyriou, Stéphan Cléménçon, and Ruofeng Zhang. Learning the graph of relations among multiple tasks. *ICML workshop on New Learning Frameworks and Models for Big Data*, 2014.
- [4] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.
- [5] Andreas Argyriou, Andreas Maurer, and Massimiliano Pontil. An algorithm for transfer learning in a heterogeneous environment. In *Machine Learning and Knowledge Discovery in Databases*, pages 71–85. Springer, 2008.
- [6] Andreas Argyriou, Massimiliano Pontil, Yiming Ying, and Charles A Micchelli. A spectral regularization framework for multi-task structure learning. In *Advances in neural information processing systems*, pages 25–32, 2007.

- [7] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, pages 1497–1537, 2005.
- [8] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, March 2003. Available from: <http://dl.acm.org/citation.cfm?id=944919.944944>.
- [9] Jonathan Baxter. A model of inductive bias learning. *J. Artif. Intell. Res.(JAIR)*, 12(149-198):3, 2000.
- [10] Shai Ben-David and Reba Schuller Borbely. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine learning*, 73(3):273–287, 2008.
- [11] Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer, 2003.
- [12] Steffen Bickel, Jasmina Bogojeska, Thomas Lengauer, and Tobias Scheffer. Multi-task learning for hiv therapy screening. In *Proceedings of the 25th international conference on Machine learning*, pages 56–63. ACM, 2008.
- [13] Olivier Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.
- [14] Olivier Bousquet. Concentration inequalities for sub-additive functions using the entropy method. In Evariste Gin, Christian Houdr, and David Nualart, editors, *Stochastic Inequalities and Applications*, volume 56 of *Progress in Probability*, pages 213–247. Birkhuser Basel, 2003. Available from: http://dx.doi.org/10.1007/978-3-0348-8069-5_14, doi:10.1007/978-3-0348-8069-5_14.
- [15] Bin Cao, Nathan N Liu, and Qiang Yang. Transfer learning for collective link prediction in multiple heterogenous domains. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 159–166, 2010.
- [16] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
- [17] Corinna Cortes, Marius Kloft, and Mehryar Mohri. Learning kernels using local rademacher complexity. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2760–2768. Curran Associates, Inc., 2013. Available from: <http://papers.nips.cc/paper/4896-learning-kernels-using-local-rademacher-complexity.pdf>.
- [18] Corinna Cortes and Mehryar Mohri. *Algorithmic Learning Theory: 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings*, chapter Domain Adaptation in Regression, pages 308–323. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. Available from: http://dx.doi.org/10.1007/978-3-642-24412-4_25, doi:10.1007/978-3-642-24412-4_25.
- [19] Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103 – 126, 2014. Algorithmic Learning Theory. Available from: <http://www.sciencedirect.com/science/article/pii/S0304397513007184>, doi:http://dx.doi.org/10.1016/j.tcs.2013.09.027.
- [20] A Evgeniou and Massimiliano Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19:41, 2007.
- [21] Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, 13(1):1865–1890, 2012.
- [22] Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 521–528, 2011.

- [23] Marius Kloft and Gilles Blanchard. The local rademacher complexity of lp-norm multiple kernel learning. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2438–2446. Curran Associates, Inc., 2011. Available from: <http://papers.nips.cc/paper/4259-the-local-rademacher-complexity-of-lp-norm-multiple-kernel-learning>
- [24] Marius Kloft and Gilles Blanchard. On the convergence rate of lp-norm multiple kernel learning. *The Journal of Machine Learning Research*, 13(1):2465–2502, 2012.
- [25] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, 30(1):1–50, 02 2002. Available from: <http://dx.doi.org/10.1214/aos/1015362183>, doi:10.1214/aos/1015362183.
- [26] Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 12 2006. Available from: <http://dx.doi.org/10.1214/009053606000001019>, doi:10.1214/009053606000001019.
- [27] Vladimir Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *J. Mach. Learn. Res.*, 11:2457–2485, December 2010. Available from: <http://dl.acm.org/citation.cfm?id=1756006.1953014>.
- [28] Abhishek Kumar and Hal Daume III. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*, 2012.
- [29] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.
- [30] Yunwen Lei, Lixin Ding, and Yingzhou Bi. Local rademacher complexity bounds based on covering numbers. *arXiv:1510.01463 [cs.AI]*, 2015.
- [31] Cong Li, Michael Georgiopoulos, and Georgios C Anagnostopoulos. Multitask classification hypothesis space with improved generalization bounds. *Neural Networks and Learning Systems, IEEE Transactions on*, 2013.
- [32] K Lounici, M Pontil, AB Tsybakov, and SA Van De Geer. Taking advantage of sparsity in multi-task learning. In *COLT 2009-The 22nd Conference on Learning Theory*, 2009.
- [33] F. Lust-Piquard. Khintchine inequalities in cp ($1 < p < \infty$). *COMPTES RENDUS DE L ACADEMIE DES SCIENCES SERIE I-MATHEMATIQUE*, 303(7):289–292, 1986.
- [34] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of The 22nd Annual Conference on Learning Theory (COLT 2009)*. Omnipress, June 2009.
- [35] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1041–1048. Curran Associates, Inc., 2009. Available from: <http://papers.nips.cc/paper/3550-domain-adaptation-with-multiple-sources.pdf>.
- [36] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the rÉnyi divergence. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 367–374, Arlington, Virginia, United States, 2009. AUAI Press. Available from: <http://dl.acm.org/citation.cfm?id=1795114.1795157>.
- [37] Yishay Mansour and Mariano Schain. Robust domain adaptation. *Annals of Mathematics and Artificial Intelligence*, 71(4):365–380, 2013. Available from: <http://dx.doi.org/10.1007/s10472-013-9391-5>, doi:10.1007/s10472-013-9391-5.
- [38] Andreas Maurer. Bounds for linear multi-task learning. *The Journal of Machine Learning Research*, 7:117–139, 2006.

- [39] Andreas Maurer. The rademacher complexity of linear transformation classes. In *Learning Theory*, pages 65–78. Springer, 2006.
- [40] Andreas Maurer. *Algorithmic Learning Theory: 25th International Conference, ALT 2014, Bled, Slovenia, October 8-10, 2014. Proceedings*, chapter A Chain Rule for the Expected Suprema of Gaussian Processes, pages 245–259. Springer International Publishing, Cham, 2014. Available from: http://dx.doi.org/10.1007/978-3-319-11662-4_18, doi:10.1007/978-3-319-11662-4_18.
- [41] Andreas Maurer and Massimiliano Pontil. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory*, volume 30, pages 55–76, 2013.
- [42] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *arXiv preprint arXiv:1505.06279*, 2015.
- [43] Shahar Mendelson. On the performance of kernel classes. *The Journal of Machine Learning Research*, 4:759–771, 2003.
- [44] Luca Oneto, Alessandro Ghio, Sandro Ridella, and Davide Anguita. Local rademacher complexity: Sharper risk bounds with and without unlabeled samples. *Neural Networks*, 65:115 – 125, 2015. Available from: <http://www.sciencedirect.com/science/article/pii/S0893608015000404>, doi:<http://dx.doi.org/10.1016/j.neunet.2015.02.006>.
- [45] Anastasia Pentina and Shai Ben-David. Multi-task and lifelong learning of kernels. In *Algorithmic Learning Theory*, pages 194–208. Springer, 2015.
- [46] Anastasia Pentina and Christoph H Lampert. Lifelong learning with non-iid tasks. In *Advances in Neural Information Processing Systems*, pages 1540–1548, 2015.
- [47] G Peshkir and Albert Nikolaevich Shiryaev. The khintchine inequalities and martingale expanding sphere of their action. *Russian Mathematical Surveys*, 50(5):849–904, 1995.
- [48] Ting Kei Pong, Paul Tseng, Shuiwang Ji, and Jieping Ye. Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM Journal on Optimization*, 20(6):3465–3489, 2010.
- [49] Bernardino Romera-Paredes, Andreas Argyriou, Nadia Berthouze, and Massimiliano Pontil. Exploiting unrelated tasks in multi-task learning. In *International Conference on Artificial Intelligence and Statistics*, pages 951–959, 2012.
- [50] Michel Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.
- [51] S Thrun. Learning to learn: Introduction. In *In Learning To Learn*, 1996.
- [52] I. Tolstikhin, G. Blanchard, and M. Kloft. Localized complexities for transductive learning. In *Proceedings of the 27th Conference on Learning Theory*, volume 35, pages 857–884. JMLR, 2014. Available from: <http://jmlr.org/proceedings/papers/v35/tolstikhin14.pdf>.
- [53] Qian Xu, Sinno Jialin Pan, Hannah Hong Xue, and Qiang Yang. Multitask learning for protein sub-cellular location prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(3):748–759, 2011.
- [54] Chao Zhang, Lei Zhang, and Jieping Ye. Generalization bounds for domain adaptation. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3320–3328. Curran Associates, Inc., 2012. Available from: <http://papers.nips.cc/paper/4684-generalization-bounds-for-domain-adaptation.pdf>.
- [55] Yu Zhang and Dit-Yan Yeung. Multi-task warped gaussian process for personalized age estimation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2622–2629. IEEE, 2010.